STIL 2026

XVII Brazilian Symposium in Information and Human Language Technology

Proceedings of the Conference, Vol. 1

October 19-22, 2026

©2026 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

About the symposium

The Proceedings of the XVII Brazilian Symposium in Information and Human Language Technology (STIL 2026) bring together the papers selected and presented at the 17th edition of the symposium, held from October 19 to 22, 2026, in Cuiabá, Mato Grosso, Brazil, in conjunction with the XXXVI Brazilian Conference on Intelligent Systems (BRACIS 2026), the XXIII National Meeting on Artificial and Computational Intelligence (ENIAC 2026), the XIV Symposium on Knowledge Discovery, Mining and Learning (KDMiLe 2026), the XX Workshop-School on Agents, Environments, and Applications (WESAAC 2026), and the XV Thesis and Dissertation Contest in Artificial and Computational Intelligence (CTDIAC 2026).

This edition of the proceedings includes papers selected for the Main Conference and for its satellite events:

- The XVII Brazilian Symposium in Information and Human Language Technology (STIL 2026) received 89 submissions, of which 37 were accepted (41.6%). Final versions of 36 papers were submitted for publication in these proceedings. The papers were selected through a double-blind peer-review process. The Program Committee was chaired by Aline Paes (UFF) and Ariani Di Felippo (UFSCar), who organized this volume together with Bryan Khelven da Silva Barbosa (UFSCar), Publication Chair of STIL 2026. The submission deadline was May 15, 2026; acceptance notifications were sent on July 11, 2026; and final versions were due on July 26, 2026.

- The XI Workshop on Portuguese Description (JDP 2026) received 13 submissions, of which 8 were accepted (61.5%); all 8 final versions are included in these proceedings. The papers were selected through a double-blind peer-review process. The event was chaired by Daniela de Souza Silva Costa (UFMS) and Maucha Andrade Gamonal (UFJF). The submission deadline was June 1, 2026; acceptance notifications were sent on July 11, 2026; and final versions were due on July 15, 2026.

- The XI Undergraduate Research Workshop on Information and Human Language Technology (TILic 2026) received 20 submissions, of which 18 were accepted (90%). Final versions of 17 papers were submitted for publication in these proceedings. The papers were selected through a double-blind peer-review process. The event was chaired by Lilian Teixeira (UFBA) and Paula Cardoso (UFPA). The submission deadline was June 5, 2026; acceptance notifications were sent on July 11, 2026; and final versions were due on August 3, 2026.

Acknowledgments

The chairs of STIL 2026 and its satellite events acknowledge the support to the conference provided by the Brazilian Computer Society (SBC). We thank the reviewers of the XVII Brazilian Symposium on Information and Human Language Technology and its satellite events for their careful reviews. Last but not least, we are grateful to Thiago Ventura (UFMT) and the BRACIS 2026 organizing team in Cuiabá for making the joint event possible.

October 2026

Aline Paes (UFF, Brazil)
Ariani Di Felippo (UFSCar, Brazil)

Organizing Committee

Program Chairs - STIL

Aline Paes, UFF, Brazil
Ariani Di Felippo, UFSCar, Brazil

Program Chairs - JDP

Daniela de Souza Silva Costa, UFMS, Brazil
Maucha Andrade Gamonal, UFJF, Brazil

Program Chairs - TILiC

Lilian Teixeira, UFBA, Brazil
Paula Christina Figueira Cardoso, UFPA, Brazil

Publication Chair

Bryan Khelven da Silva Barbosa, UFSCar, Brazil

Proceedings Organization

Bryan Khelven da Silva Barbosa, UFSCar, Brazil
Aline Paes, UFF, Brazil
Ariani Di Felippo, UFSCar, Brazil

Program Committee

Technical Program Committee

Adriana Pagano, Universidade Federal de Minas Gerais (UFMG)
Anderson Castro Soares de Oliveira, Universidade Federal de Mato Grosso (UFMT)
André Carvalho, Universidade Federal do Amazonas (UFAM)
Arnaldo Candido Junior, Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP)
Bryan Khelven da Silva Barbosa, Universidade Federal de São Carlos (UFSCar)
Carlo Marcelo Revoredo da Silva, Universidade de Pernambuco (UPE)
Carlos Alberto Severiano Júnior, Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais (IFMG)
Carlos Augusto Prolo, Universidade Federal do Rio Grande do Norte (UFRN)
Carlos Caminha, Universidade Federal do Ceará (UFC)
Cassio Faria Silva, Faculdade de Tecnologia de Marília (FATEC Marília)
Christopher Dane Shulby, Universidade de São Paulo (USP)
Cláudia Bosaipo Sales da Silva, Instituto Militar de Engenharia (IME)
Claudia Dias de Barros, Universidade Federal de São Carlos (UFSCar)
Cláudio Elízio Calazans Campelo, Universidade Federal de Campina Grande (UFCG)
Cleyton Mário de Oliveira Rodrigues, Universidade de Pernambuco (UPE)
Daniela Barreiro Claro, Universidade Federal da Bahia (UFBA)
Denis Deratani Mauá, Universidade de São Paulo (USP)
Dennis Balreira, Universidade Federal do Rio Grande do Sul (UFRGS)
Diana Santos, Linguateca / Universidade de Oslo (UiO)
Eduardo Bezerra, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ)
Eduardo C. Goncalves, Escola Nacional de Ciências Estatísticas (ENCE)
Eduardo Jose da Silva Luz, Universidade Federal de Ouro Preto (UFOP)
Elisa Marchioro Stumpf, Universidade Federal do Rio Grande do Sul (UFRGS)
Eloize Rossi Marques Seno, Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP)
Emerson Cabrera Paraiso, Pontifícia Universidade Católica do Paraná (PUCPR)
Evelin Amorim, Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência (INESC TEC)
Fábio Lobato, Universidade Federal do Oeste do Pará (UFOPA)
Gabriela Wick Pedro, Universidade Federal de São Carlos (UFSCar)
Gustavo Guedes, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ)
Gustavo Zanatta Bruno, Universidade do Vale do Rio dos Sinos (UNISINOS)
Helena Cameron, Centro Interdisciplinar de História, Culturas e Sociedades / Universidade de Évora (CIDEHUS/UÉ)
Helena de Medeiros Caseli, Universidade Federal de São Carlos (UFSCar)
Helen Cristina de Mattos Senefonte, Universidade Estadual de Londrina (UEL)
Hilário Tomaz Alves de Oliveira, Instituto Federal de Educação, Ciência e Tecnologia do Espírito Santo (IFES)
Hugo Gonçalo Oliveira, Universidade de Coimbra (UC)
Igor Cataneo Silveira, Instituto de Matemática e Estatística da Universidade de São Paulo (IME-USP)
Isabel Harb Manssour, Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)
Ivone Penque Matsuno, Universidade Federal de Mato Grosso do Sul (UFMS)
Jackson Wilke da Cruz Souza, Universidade Federal da Bahia (UFBA)
Jayr Pereira, Universidade Federal do Cariri (UFCA)
Joao Paulo Papa, Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP)
Joel Luís Carbonera, Universidade Federal do Rio Grande do Sul (UFRGS)
Jonnathan Carvalho, Instituto Federal de Educação, Ciência e Tecnologia Fluminense (IFF)
Jorge Baptista, Universidade do Algarve (UAlg)
José Antonio Macêdo, Universidade Federal do Ceará (UFC)
José Wellington Franco da Silva, Universidade Federal do Ceará (UFC)
Julio Cesar Soares dos Reis, Universidade Federal de Viçosa (UFV)
Juscimara Avelino, Universidade Federal de Pernambuco (UFPE)
Leandro Krug Wives, Universidade Federal do Rio Grande do Sul (UFRGS)

Leandro Passos, Universidade Estadual Paulista “Júlio de Mesquita Filho” (UNESP)
Leonardo Rocha, Universidade Federal de São João del-Rei (UFSJ)
Leonel Figueiredo de Alencar, Universidade Federal do Ceará (UFC)
Lilian Berton, Universidade Federal de São Paulo (UNIFESP)
Lilian Sousa, Universidade Federal da Bahia (UFBA)
Livy Maria Real Coelho, Instituto de Computação da Universidade Federal do Amazonas (IComp/UFAM)
Lucas Weihmann, Universidade Federal de Santa Catarina (UFSC)
Lucelene Lopes, Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (ICMC-USP)
Luciana Borges, Universidade Federal de Mato Grosso (UFMT)
Marcos Fernando Lopes, Universidade de São Paulo (USP)
Maria das Graças Volpe Nunes, Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (ICMC-USP)
Maria José Bocorny Finatto, Programa de Pós-Graduação em Letras da Universidade Federal do Rio Grande do Sul (PPGLet/UFRGS)
Marlo Vieira dos Santos e Souza, Universidade Federal da Bahia (UFBA)
Matheus Bernardelli de Moraes, Centro de Tecnologia da Informação Renato Archer (CTI Renato Archer)
Maucha Andrade Gamonal, Universidade Federal de Juiz de Fora (UFJF)
Miguel Arjona Ramírez, Universidade de São Paulo (USP)
Miguel Oliveira Jr., Universidade Federal de Alagoas (UFAL)
Myriam Regattieri De Biase da Silva Delgado, Universidade Tecnológica Federal do Paraná (UTFPR)
Norton Trevisan Roman, Escola de Artes, Ciências e Humanidades da Universidade de São Paulo (EACH-USP)
Omar Andres Carmona Cortes, Instituto Federal de Educação, Ciência e Tecnologia do Maranhão (IFMA)
Oto Araújo Vale, Universidade Federal de São Carlos (UFSCar)
Pablo Andretta Jaskowiak, Universidade Federal de Santa Catarina (UFSC)
Pablo Gamallo, Universidade de Santiago de Compostela (USC)
Patricia Medyna Lauritzen de Lucena Drumond, Universidade Federal do Piauí (UFPI)
Paula Christina Figueira Cardoso, Universidade Federal do Pará (UFPA)
Paulo Mann, Universidade Federal do Rio de Janeiro (UFRJ)
Rafael Anchiêta, Instituto Federal de Educação, Ciência e Tecnologia do Maranhão – Campus Caxias (IFMA – Campus Caxias)
Raphael de Souza Rosa Gomes, Universidade Federal de Mato Grosso (UFMT)
Raquel Meister Ko. Freitag, Universidade Federal de Sergipe (UFS)
Renata Vieira, Universidade de Évora (UÉ)
Renato Fileto, Universidade Federal de Santa Catarina (UFSC)
Renato Moraes Silva, Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (ICMC-USP)
Ricardo Marcondes Marcacini, Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (ICMC-USP)
Roana Rodrigues, Universidade Federal de Sergipe (UFS)
Robson Fidalgo, Universidade Federal de Pernambuco (UFPE)
Rodrigo Minetto, Universidade Estadual de Campinas (UNICAMP)
Roney Lira de Sales Santos, Universidade Federal da Bahia (UFBA)
Saullo Haniell Galvão de Oliveira, Escola de Matemática Aplicada da Fundação Getulio Vargas (FGV EMAp)
Sergio Antonio Andrade de Freitas, Universidade de Brasília (UnB)
Thiago Alexandre Salgueiro Pardo, Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (ICMC-USP)
Thiago Ferreira, Universidade Federal Fluminense (UFF)
Thiago Henrique Silva, Universidade Tecnológica Federal do Paraná (UTFPR)
Thiago Meirelles Ventura, Universidade Federal de Mato Grosso (UFMT)
Valéria Delisandra Feltrim, Universidade Estadual de Maringá (UEM)
Valeria de Paiva, Nuance Communications, Inc. (Nuance)
Viviane Pereira Moreira, Universidade Federal do Rio Grande do Sul (UFRGS)
Vladia Pinheiro, Universidade de Fortaleza (UNIFOR)

Additional Reviewers

Felipe Ribas Serras, Instituto de Matemática e Estatística da Universidade de São Paulo (IME-USP)
João Pedro Tomaszewski, Universidade Federal de Pelotas (UFPel)
João Vítor Vaz, Universidade Federal de Minas Gerais (UFMG)
Matheus Yasuo Ribeiro Utino, Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (ICMC-USP)
Wesley Carvalho, Universidade de São Paulo (USP)

Table of Contents

Extending Enhanced Universal Dependencies to Nheengatu: Semi-Automated Control and Raising Annotation
Dominick Maia Alexandre and Leonel Figueiredo de Alencar 1

Leakage-Aware Evaluation of Brazilian Clinical Notes for ICU Mortality Prediction: A Study on the BRATECA Dataset
Felipe André Bach Alves, Heloísa Oss Boll and Mariana Recamonde-Mendoza 14

Do LLM Judges Agree with Humans? Evaluating Financial Commentaries from Material Facts
Gabriel Assis, Daniela Vianna, Livy Real, Marina Ramalhete Masid, João Nepomuceno, Eduardo Rottschaefer, Altigran Soares da Silva and Aline Paes 26

Expressões multipalavras no discurso de fóruns sobre saúde mental: desenvolvimento de um corpus em inglês e português brasileiro
Gabriela Berndt, Sofia Serrano, Mariana Jin and Adriana Pagano 41

Text segmentation with mixed data: evaluating symbolic and neural approaches
Amanda Sammer do N. Brandão, Felipe Bomfim Sanches, Alfredo Sena, Laila Mota, Daniela Barreiro Claro and Rerisson Cavalcante 50

Exploring Hybrid Pre-training for Automatic Essay Scoring
Miguel M. Carpi, Igor C. Silveira, Denis D. Mauá and Marcelo Finger 62

ENEM Essay Feedback Using LLM-Augmented Prompts
Flavio Carvalho, Vanessa Soares, Eduardo Bezerra and Gustavo Guedes 78

Empirical Study Towards RBAMR-News: a Rule-Based AMR Parser for News in Portuguese
Maria Julia Bernardo Comarim and Ariani Di-Felippo 86

A Comparative Framework for Diachronic Lexical Semantic Change in Brazilian Portuguese Political Discourse
João Victor M. Correa, Oto Araújo Vale and Renato Moraes Silva 100

Autonomous Fact-Checking: Integrating Local Inference and Adversarial NLP Attacks
Rômulo Henrique Nascimento Duarte, Humberto Nunes de Lira, Vivianny Khatly Medeiros Pereira, Isaac Sebastian Lima de Araújo, Allan Gabriel da Cunha Vasconcelos and Yuri de Almeida Malheiros Barbosa 113

A Comparison of Commonly Used Automatic Evaluation Metrics for Open-Ended Tasks in Portuguese
Eduardo D. Faé, Dennis Giovani Balreira and Viviane Pereira Moreira 126

DANTE – A Treebank of Tweets in Brazilian Portuguese
Ariani Di-Felippo and Norton Trevisan Roman 139

EpiCorpus-BR: A Named Entity Recognition Corpus for Epidemiological Documents in Portuguese
Christian Freitas and Lilian Berton 150

Curupira: Arquiteturas Híbridas Transformer-BiLSTM para Detecção de Discurso Ofensivo em Português Brasileiro
Gabrielly Alves Gomes, Iago de Sousa Aragão and Patricia Medyna Lauritzen de Lucena Drumond 163

CorPop26: Uma atualização das 3.000 Palavras Mais Frequentes do CorPop de 2018
Roges Horacio Grandi, Bianca Pasqualini, Maria Jose Bocorny Finatto, Leandro Krug Wives and Raquel Salcedo Gomes 175

Vidya Syl: A syllable division algorithm for Brazilian Portuguese adhering to the Common Orthographic Vocabulary of the Portuguese Language
Roges Horacio Grandi, Leandro Krug Wives, Maria Jose Bocorny Finatto, Raquel Salcedo Gomes and Bianca Pasqualini 188

Developing Multilingual Multiword Expressions (MWEs) Resources: An Annotated Corpus of Fables following the PARSEME guidelines
Letícia Guedes Guimarães, Adriana Pagano, André V. Lopes Coneglian and Aline Villavicencio 200

Beyond Aggregate Scores: A Diagnostic Analysis of Portuguese LLM Evaluation Suites
Gustavo Hochgraf, Gabriel Assis, Aline Paes and Fabio G. Cozman 209

Balancing Coherence and Granularity in Legal Topic Modeling: A BERTopic-based Study on Portuguese Appellate Court Decisions
Marjory S. S. Lima, Felipe R. Ahad, Vih A. S. Silva, Thiago M. Ventura, Josiel M. Figueiredo and Allan G. de Oliveira 218

Bootstrapping Text Anomaly Detection with LLM-Generated Weak Supervision
Fabio Masaracchia Maia and Anna Helena Reali Costa 231

Evaluating Portuguese Tokenizers as Morpheme Sequence in Relation to LLM Downstream Performance
Guilherme L. Mello and Marcelo Finger 245

Tail Smoothing and Cross-Lingual Volatility: Evaluating Estimative Uncertainty in Large Language Models for Brazilian Portuguese
Renato O. Miyaji and Pedro L. P. Corrêa 258

Textbook-Enriched Training for Language Models: Boosting Answer Quality in Specialized Contexts
Lucas B. Bulcão Mota, Larrissa Dantas, Daniela Barreiro Claro, Aline Paes, Claudia Freitas, Marlo Souza, Helena Caseli and Livy Real 270

Towards Uniform Meaning Representation for Brazilian Portuguese: Building a First UMR-Annotated Dataset
Adriana S. Pagano, Magali Duran, Federica Gamba and Daniel Zeman 285

Novo CorPop Saúde: Data Collection for Simplified Health Communication in a Brazilian Patient Information Leaflet Corpus
Bianca Pasqualini, Maria J. B. Finatto, Eduarda Bertotto, Paula Salem Carpio, Roges Horacio Grandi and Leandro Krug Wives 297

Legal Nugget Extraction for Granular Retrieval over Long Jurisprudential Texts
Lucas Pereira, Erick Brito, Roberto Lotufo and Jayr Pereira 305

The Impact of Language and Portuguese Intralinguistic Variation on NL2SQL
Ricardo Trainotti Rabonato and Lilian Berton 317

Como Nossos Pais: Princípios Clássicos para a Avaliação de Sistemas de IA
Livy Real and Altigran Soares da Silva 331

Clonagem de Voz por IA para o Português Brasileiro em Contexto Assistivo: Desafios Comparativos em Cinco Ferramentas de Síntese de Fala
Luiz G. dos Santos, Nathalia Borges, Matheus Sales, Luciana C. L. F. Borges, Thais R. Kempner and Eunice P. S. Nunes 348

Evaluating LLM-based Triple Extraction for Knowledge Graph Fact-Checking in Portuguese
Roney Lira de Sales Santos, Lucas dos Santos and João Pedro Holanda Souza 361

Prompt Engineering for Small Language Models: Evaluating ICL for Portuguese Sentiment Analysis
André da F. Schuck, Gabriel L. Garcia, João Renato R. Manesco, Pedro Henrique Paiola, Leandro A. Passos and João Paulo Papa 375

Compression-Based Linguistic Complexity Metrics in Automatic Essay Scoring
Felipe Ribas Serras, Igor Cataneo Silveira, Denis Deratani Mauá and Marcelo Finger 390

Validating Semantic Content in Sentence Simplification using OpenIE and Relational Graphs
Samuel Rios da Silva, Larrissa Dantas, Daniela Barreiro Claro, Aline Paes, Maria José Finatto, Rerisson Cavalcante and Silvana Ribeiro 405

Fusão Híbrida para Recuperação de Informação Jurídica Brasileira
Flávio Soriano, Guilherme Evangelista, Celso França, Gisele Pappa, Wagner Meira Jr. and Marcos Gonçalves 420

Sector-Specific Financial Sentiment Lexicons for Portuguese: Evidence from Brazilian Stock News
Antônio Artur de Souza, Mateus Binda and Adriana S. Pagano 433

MultiDomainYT-ToxBR: Um Corpus Multidomínio de Comentários Tóxicos do YouTube em Português Brasileiro
Gabriel Z. de Souza and Isabel H. Manssour 441

An Exploratory Study of Discursive Signals Annotation in User-Generated Content
Rivânia da Paixão de Jesus Carvalho, Mateus Araújo Pereira and Jackson Wilke da Cruz Souza 454

Do pequeno Asteróide B-612 ao universo: a anotação de O Pequeno Príncipe segundo o modelo Universal Dependencies
Magali Sanches Duran, Lucelene Lopes, Maria das Graças Volpe Nunes and Thiago Alexandre Salgueiro Pardo 464

Classificação automática de sentenças faladas com arquitetura híbrida: análise de estruturas topicalizadas, anti-topicalizadas e canônicas
Gerson Vitor P. Fernandes, Vitoria Maria A. Silva and Cid Ivan C. Carvalho 476

A Comparative Analysis of Adjectives in Image Descriptions in Brazilian Portuguese and English
Maucha Andrade Gamonal, Adriana Pagano, Felix Engler, André V. Lopes Coneglian and Ely E. S. Mattos 487

Clause Linkage in Brazilian Portuguese Narratives: Evidence from Enhanced Dependency Annotation
Letícia Guedes Guimarães, Adriana Pagano, André V. Lopes Coneglian and Magali Duran . . 499

Inferência Textual Zero-Shot em Português: Especialização Linguística vs. Raciocínio Explícito em Modelos de Linguagem Compactos
Gabriel Silvestre Mancini and Marcos Lopes 512

Quando o "..." importa: análise comparada de modalidades de transcrição de hesitações em entrevistas do Corpus Roda Viva
Valeria Vieira dos Santos 522

Da sintaxe ao discurso: Evidências introdutórias da correlação entre Universal Dependencies e Rhetorical Structure Theory
Jackson Wilke da Cruz Souza and Paula Christina Figueira Cardoso 531

Classificação de Textos com Matrizes de Similaridade: Um Estudo de Caso em um Corpus com Nomes de Produtos
Antonio Ferreira de Castro Barbosa and Eduardo Corrêa Gonçalves 540

Uso de modelos de linguagem na identificação de relações de coerência da Rhetorical Structure Theory em textos de User-Generated Content
Pedro Moreno Niella de Souza Sales Evangelista and Jackson Wilke da Cruz Souza 546

Expansão da Ferramenta Doccano para a Anotação de Análise de Sentimento Estruturada
Antonio Paulo Martins Ferri, Guilherme Dallmann Lima, Larissa Astrogildo de Freitas and Andrew Campos 551

A Comparative Study of Discourse Signal Proposals in Rhetorical Structure Theory
Gabriela Fortalesa and Jackson Wilke da Cruz Souza 556

Identificação e classificação de trechos racistas e sexistas em processos judiciais
Yasmin Francisquetti Barnes and Marcelo Finger 562

Data-Faithful Natural Language Generation for Explaining Fitness-Market Anomalies
Kauan Divino Pouso Mariano, Fabrycio Leite Nakano Almada, Victor Emanuel da Silva Monteiro, Maykon Adriell Dutra, Jefferson Felex de Faria, Jennifer Vitória da Silva Peixoto and Kamylla Sejane Pouso Freitas 567

OABench-Consumidor: Um benchmark para avaliação de modelos de linguagem em Direito do Consumidor
Lucas B. Bulcão Mota, Aline Athaydes, Babacar Mane and Daniela Barreiro Claro 573

Análise de comportamento do modelo NotebookLM em transcrição automática
Brenda R. F. de Oliveira and Oto Araújo Vale 578

Desafios Tipológicos na Construção de Treebanks: O Mapeamento Morfossintático da Língua Kubeo no Padrão UD
Pedro Paulo Freire Oliveira and Thiago Blanch Pires 583

Unsupervised Subword Segmentation for POS Tagging in Low-Resource Agglutinative Languages
Jonas Oliveira Pereira, Lilian Teixeira de Sousa and Marlo Souza 588

Explorando a Integração de Múltiplas Modalidades Visuais no Reconhecimento Contínuo de Libras
Guilherme Galvão de Oliveira Pinto, Guilherme Vieira Leite, Hélio Pedrini and José Mario De Martino 594

Da Definição das Entidades à Anotação do Corpus: Desafios da tarefa no Direito do Consumidor
Samuel Rios da Silva, Aline Athaydes, Babacar Mane and Daniela Barreiro Claro 600

Anotação de Metáforas em Saúde Mental: Análise de Postagens Baseada no Procedimento MIPVU
Isabela C. Rodrigues and Helena de Medeiros Caseli 605

Esquecida no Churrasco: Um método para validação e correção responsável de traduções automáticas do BBQ para Português do Brasil
Luiza Rodrigues Cardoso and Leo Sampaio Ferraz Ribeiro 610

Integração de Conhecimento Linguístico Estruturado em LLMs para a Tarefa de Sumarização Extrativa
Carlos Vitor Cardoso da Silva and Paula Christina Figueira Cardoso 616

Concatenative Synthesis for Brazilian Sign Language
Gustavo Borba da Silva, Ângelo Brandão Benetti and José Mario De Martino 621

Heurísticas vs. LLMs na triagem de tweets: avaliando pistas linguísticas para tarefas de PLN em português
Laura Pessine Teixeira and Helena de Medeiros Caseli 626

Extending Enhanced Universal Dependencies to Nheengatu: Semi-Automated Control and Raising Annotation

Dominick Maia Alexandre[1], **Leonel Figueiredo de Alencar**[1,2]

[1]Centro de Humanidades – Universidade Federal do Ceará (UFC)
Av. da Universidade, 2683 – 60020-181 – Fortaleza – CE – Brazil

[2]Instituto de Estudos da Linguagem – Universidade Estadual de Campinas (Unicamp)
R. Sérgio Buarque de Holanda, 571 – 13083-859 – Campinas – SP – Brazil

dominick@letras.ufc.br, leonel.de.alencar@ufc.br

Abstract. *This paper presents the first Enhanced Universal Dependencies annotation effort for a Brazilian Indigenous language, focusing on XCOMP constructions in Nheengatu, an endangered Tupian language. We describe a semi-automatic pipeline for the UD_Nheengatu-CompLin treebank to annotate control and raising predicates and propagated subjects. The method combines theoretical discussions on XCOMP constructions with corpus-based analysis of Nheengatu syntax to create rule-based enhancement heuristics. The work also provides annotation guidelines and expands EUD resources for Indigenous and low-resource languages within the Universal Dependencies framework.*

1. Introduction

Universal Dependencies (UD) has become a widely adopted framework for cross-linguistic morphosyntactic annotation, encompassing, as of release 2.17, 339 treebanks representing 186 languages [Nivre et al. 2016, Church and Liberman 2021, Zeman et al. 2025]. Beyond basic dependency trees, the framework defines an optional enhanced annotation layer, known as Enhanced Universal Dependencies (EUD), which is designed to make implicit syntactic relations explicit in dependency graphs [Schuster and Manning 2016].

EUD graphs cover 6 types of enhancements: (i) empty nodes for elided predicates; (ii) propagation of incoming dependencies to conjuncts; (iii) propagation of outgoing dependencies from conjuncts; (iv) additional subject dependencies for control and raising constructions; (v) coreference in relative clause constructions; and (vi) case or mark information for adpositions and conjunctions [Universal Dependencies 2026].

Although EUD annotation has expanded for high-resource languages such as Portuguese [Pagano et al. 2023, Souza et al. 2024, Sanches Duran et al. 2025], it remains uncommon in low-resource UD treebanks. In this context, Nheengatu, an endangered Tupian language spoken mainly in the Upper Rio Negro region of Brazil [Navarro 2012], is represented by UD_Nheengatu-CompLin, the only syntactically annotated corpus of the language within the UD framework [de Alencar 2023, de Alencar 2024, de Alencar 2025, Alencar et al. 2025]. Until now, however, the treebank has lacked guidelines for EUD annotation.

To address this gap, we propose the first EUD annotation effort for a Brazilian Indigenous language, focusing on XCOMP constructions in Nheengatu. Several morphosyntactic phenomena in the language pose challenges for computational modeling,

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 1–13
October 19-22, 2026 ©2026 Association for Computational Linguistics

including pro-drop, clause chaining, and argument sharing, whose underlying syntactic relations are not recoverable from basic dependency trees alone. Moreover, unlike the canonical non-finite XCOMP patterns described in the literature, verbs heading XCOMP clauses in Nheengatu are finite and inflected for person. Our approach combines theoretical analyses of control and raising with a rule-based enhancement pipeline implemented in Python using the `conllu` library. Based on a corpus analysis of Nheengatu syntax, we formalize subject-control, object-control, raising-like configurations, and propagated subjects in the EUD layer, following the official UD guidelines.

The remainder of this paper is organized as follows. Section 2 contextualizes our work within Amerindian-language UD treebanks and prior EUD research. Section 3 discusses XCOMP constructions in Nheengatu. Section 4 presents the enhancement pipeline and heuristics. Section 5 reports the annotation results and error analysis, and Section 6 provides concluding remarks and future directions.

2. Related Work

As of the Universal Dependencies release 2.17 (November 15, 2025), the UD collection comprises 25 treebanks covering 24 Amerindian languages, including two treebanks for Mbya-Guarani [Thomas 2019]. Among these resources, UD_Nheengatu-CompLin is the second largest in terms of token count, surpassed only by UD_Bororo-BDT.

These 25 treebanks constitute only 7.4% of the 339 treebanks and less than 1% of the 37 million tokens of the UD collection [Zeman et al. 2025]. Among them, only two treebanks include EUD annotations: Western Sierra Puebla Nahuatl [Pugh et al. 2022] and Highland Puebla Nahuatl [Pugh and Tyers 2024].

Amerindian UD treebanks also vary in corpus depth (e.g., use of POS tags, features, genre, size), as reflected in the official UD evaluation metrics. Most receive between 0 and 2.5 stars[1], indicating ongoing development and uneven syntactic coverage. The UD_Nheengatu-CompLin, however, has a rating of 3.5 stars, the highest among Amerindian-language UD treebanks. Despite its small size (2,839 sentences), its 3.5-star rating matches that of UD_Portuguese-DANTEStocks [Di Felippo et al. 2021] and approaches the 4-star rating of UD_Portuguese-Porttinari [Duran et al. 2023], both of which represent a high-resource language.

Previous work by [Guillaume and Perrier 2021] presented a rule-based graph rewriting system that converted basic UD graphs into EUD graphs for 17 languages. For five high-resource languages, the authors derived language-specific lexicons of control and raising predicates to support their annotation. Besides the fact that Nheengatu has not previously received EUD, its existing lexical resources also lack a control/raising lexicon. As EUD annotation for Nheengatu expands, using the rule-based approach of [Guillaume and Perrier 2021] may become a viable alternative, a strategy that has already proved effective for Brazilian Portuguese [Sanches Duran et al. 2025].

For Nheengatu, the need for EUD arises from morphosyntactic properties that challenge syntactic representation, particularly overt subjects in open clausal complements, which constitute the starting point of the EUD annotation proposed in this

[1]This distribution was obtained automatically through the analysis of the `eval.log` files of all 25 Amerindian-language treebanks available in UD v2.17.

work. Their annotation involves not only canonical verbal XCOMP constructions but also finite embedded predicates, a phenomenon previously discussed for Portuguese by [de Alencar and Rademaker 2022].

3. Open Clausal Complements in Nheengatu

The concept and terminology of XCOMP as an *open clausal complement* originate in Lexical-Functional Grammar (LFG), particularly in the distinction between functional and anaphoric control formalized by [Bresnan et al. 2015]. In LFG, an XCOMP is defined as an open complement that lacks an overt subject, without direct evidence for an empty NP position within the complement clause [Kroeger 2004, p. 111]. Such constructions occur only when specified in the lexical entry of the matrix predicate, whose verbs are commonly referred to as *control predicates*.

Universal Dependencies adopts this view by analyzing XCOMP constructions as subjectless core arguments in obligatory control relations, typically controlled by the matrix object or subject [de Marneffe et al. 2025].

While Head-Driven Phrase Structure Grammar (HPSG), like LFG, analyzes XCOMP as a core argument whose subject is determined externally, the two frameworks differ in how they represent control and clause relations. LFG models XCOMP through functional control in f-structure, while HPSG represents these relations through structure-sharing and valency features. HPSG also makes a distinction between control and raising constructions, whereas LFG traditionally treats both through *functional control*.

Our analysis of UD_Nheengatu-CompLin identified four main XCOMP construction types: subject-control, object-control, raising-like, and subject propagation.

3.1. Subject Control

In subject-control constructions, the understood subject of the embedded predicate is obligatorily coreferential with the matrix subject; the EUD layer makes this explicit via an enhanced `nsubj` arc from the `xcomp` head to the matrix subject node [Schuster and Manning 2016]. Three semantic classes of subject-control predicates in Nheengatu were identified: desiderative predicates, motion predicates with a purposive interpretation, and emotive predicates.

Desiderative predicates. Following [Abeillé 2024], control predicates can be classified according to the semantic class of the matrix verb, with volitional, desiderative, and evaluative verbs typically triggering subject control. Figure 1 illustrates a prototypical desiderative construction, in which the matrix subject *Adana* controls the understood subject of the embedded predicate *mendari* (‘marry’).

Motion predicates with purposive interpretation. A productive subclass of subject-control constructions involves motion predicates, represented in our corpus by *sú* (‘to go’), which accounts for 11 occurrences. In these constructions, the motion verb functions as a matrix head, selecting a dependent predicate denoting purpose, goal, or sequential action associated with the movement event.

Figure 2 illustrates this. The matrix nominal subject *Payé* (‘shaman’) controls the understood subject position of the embedded predicate *pusanú* (‘to heal’). Although both verbs bear independent person-marking prefixes ($_{3SG}$), a common pattern in Nheengatu,

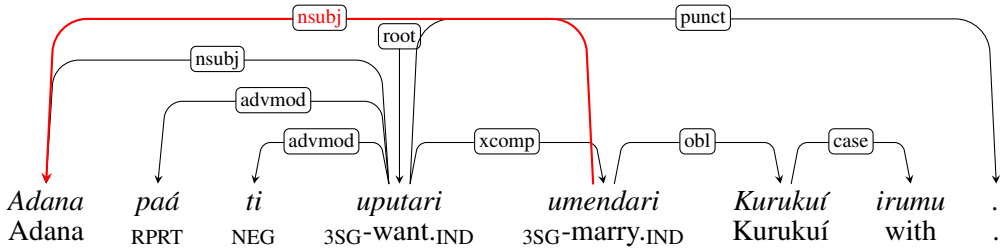


Figure 1. Enhanced dependencies for Casasnovas2006:12:15:181: 'Adana did not want to marry Kurukuí.' [Casasnovas 2006, p. 99]

they form a unified event in which the embedded predicate cannot license its own NP subject.

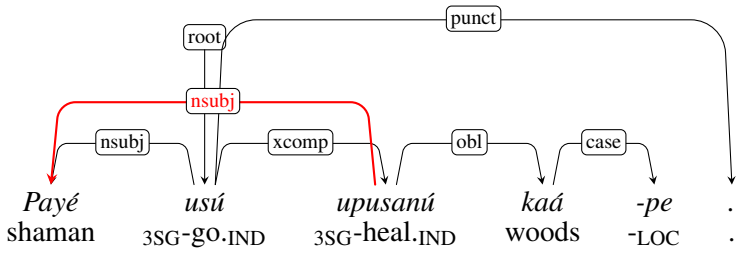


Figure 2. Enhanced dependencies for Hartt1938:0:0:509: 'The shaman went to heal in the woods'. [Hartt 1938, p. 356]

Emotive and preference predicates. A third subclass of subject control is formed by psychological and preferential predicates, exemplified by verbs such as *yumutari* ('like/enjoy'). Unlike desiderative predicates, which assign a volitional Agent role to the matrix subject, emotive predicates assign an Experiencer role: the matrix subject is an entity for whom the embedded event constitutes an object of affect or preference. Despite this semantic difference, the syntactic behavior is the same: the embedded predicate is an open clausal complement whose subject is obligatorily controlled by the matrix subject.

An example is shown in Figure 3. The matrix subject *Yawara* ('dog') functions as the Experiencer of the emotive predicate and obligatorily controls the understood subject of the embedded predicate *yenú* ('lie down').

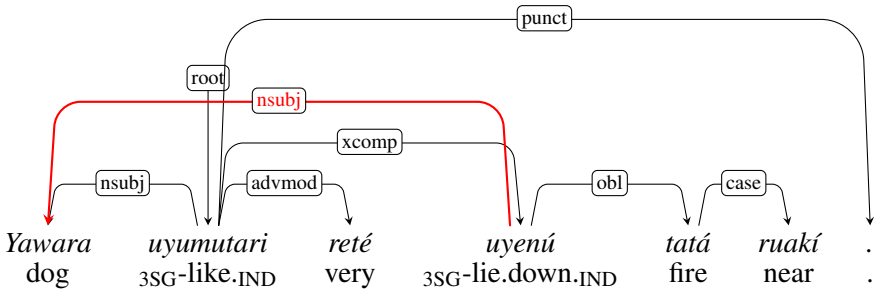


Figure 3. Enhanced dependencies for Hartt1938:0:0:141: 'The dog very much likes to lie down near the fire.' [Hartt 1938, p. 329]

Subject control also extends beyond canonical verbal predicates to inchoative constructions. The stative predicate *pitá* ('to become') selects adjectival or nominal

XCOMP complements controlled by the matrix subject, as in *upitá sasiára* (‘[He] was sad’).[2] When the matrix subject is phonologically null (pro-drop), no enhanced `nsubj` arc can be instantiated in the EUD layer, since the controller is recoverable only through verbal agreement morphology.

3.2. Object Control

In object-control constructions, the understood subject of the embedded predicate is coreferential with the matrix object. In the EUD representation, the same nominal node receives both a basic `obj` relation and an enhanced `nsubj` relation. Three main semantic subclasses of object-control predicates are attested in the corpus.

Manipulative predicates such as *purandú* (‘ask’), *mundú* (‘order/send’), and *mungitá* (‘advise’) denote events in which the matrix subject induces the matrix object to perform the embedded action. In Figure 4, *Jesus* functions both as the `obj` of *upurandú* and as the understood subject of the embedded predicate *uxari* (‘let’), formalized in the EUD layer through an enhanced `nsubj` relation.

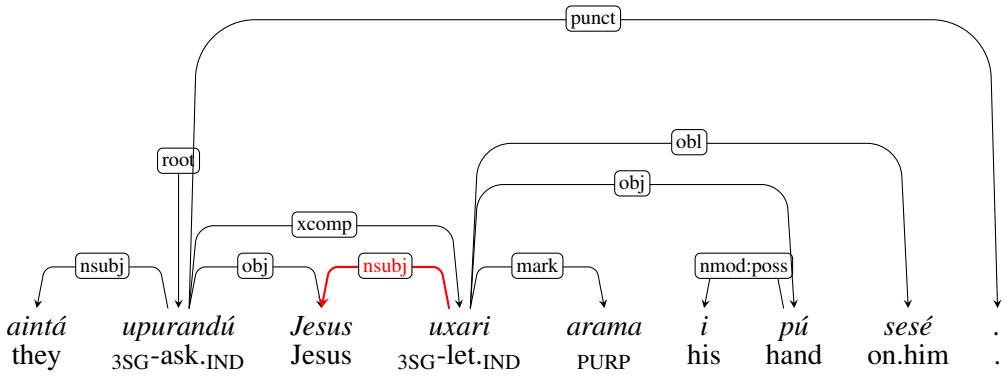


Figure 4. Enhanced dependencies for NTLN2019:5:1:37: ‘they begged Jesus to touch him.’

Perception predicates such as *maã* (‘see’), *sendú* (‘hear’), and *saã* (‘feel’) form another subclass of object-control. In these, the matrix object denotes the perceived entity while simultaneously functioning as the understood subject of the embedded predicate. This pattern is illustrated in (1), where the object *indé* (‘you’) is interpreted as the understood subject of the predicate *amurupí* (‘be different’).

(1) *Wirandé xa-maã ranhẽ indé amurupí.*
yesterday 1SG-see.IND IMPF you.ACC differently
‘Yesterday I still saw you differently.’ [de Amorim 1928, p. 99]

Naming predicates such as *musera* (‘name’), *seruka* (‘name/call’), and *senúi* (‘call’) select a matrix object corresponding to the entity being named and a nominal `xcomp` complement expressing the assigned designation. Figure 5 illustrates this pattern with the predicate *musera* (‘name’), where the object *aé* (‘it’) is interpreted as the understood subject of the nominal predicate *Carminha*.

[2] See sentence `Casasnovas2006:1:2:2` in UD_Nheengatu-CompLin.

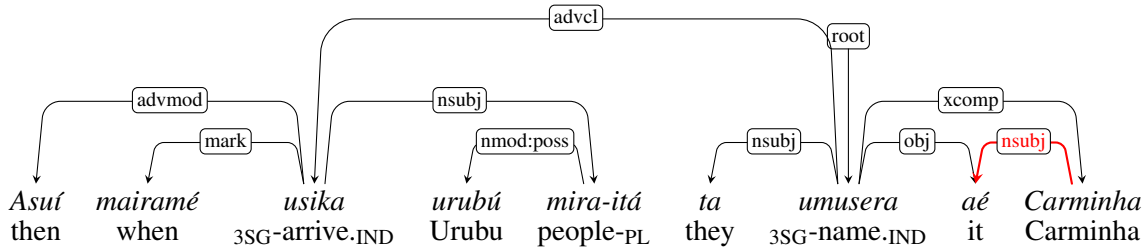


Figure 5. Enhanced dependencies for Avila2021:40:2:487: 'Then, when the people of the Urubu ethnicity arrived, they named it "Carminha".'

3.3. Raising-like predicates

Raising predicates do not assign an independent thematic role to the raised argument; its semantic role is determined by the embedded predicate [Kroeger 2004]. Although control and raising predicates are both annotated as `xcomp` in basic UD, we keep this distinction to facilitate rule refinement and future pipeline development. Since Nheengatu lacks non-finite verb forms, EUD annotation relies more heavily on lexical information.

The most frequent raising-like context is the inceptive aspectual predicate *yupirú* ('start'): the matrix subject's semantic role derives entirely from the embedded predicate, and *yupirú* itself contributes only onset aspectual information without introducing a participant role. Figure 6 illustrates this pattern.

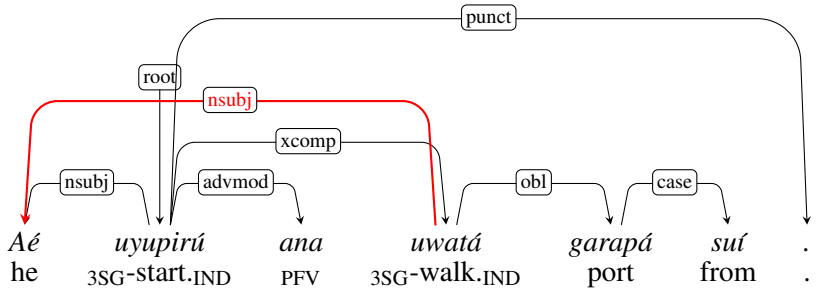


Figure 6. Enhanced dependencies for Navarro2016:0:0:150: 'He started walking from the port.'

3.4. Propagation of implicit subjects

A fourth category involves XCOMP controllers not found in the immediately local matrix domain, requiring recovery across coordination or paratactic boundaries. This pattern is frequent in the corpus: Nheengatu's use of clause-chaining via `parataxis` frequently produces chains in which a single overt subject controls embedded predicates several nodes away. Example (2) illustrates this: the noun phrase *Kurumĩ* ('boy') functions as the understood subject of the distant predicate *uyenú* ('lie down'). Making this kind of implicit argument sharing explicit is one of the enhancements the EUD layer was designed to capture [Schuster and Manning 2016].

(2) *Kurumĩ u-sú renundé u-inú suíwa rakapira upé urarí , arirí*
boy 3SG-go.IND before 3SG-put.IND his.arrow tip -LOC curare , then
u-sú u-yenú kuruatá pupé inayawasú aresé .
3SG-go.IND 3SG-lie.IND spathe inside inajá.AUG on .

'The boy first put poison on the tip of the arrow, then went to lie down on the spathe of the large inajá.' [Rodrigues 1890, p. 56]

4. Methodology

The pipeline consists of two scripts operating over CoNLL-U files: a quantitative feature-extraction script (`xcomp_context.py`) and a rule-based graph-enrichment script (`xcomp_enhancer.py`).[3]

4.1. Corpus profiling and feature extraction

We processed a pre-filtered subset containing only sentences with at least one `xcomp` relation. For each target token, the script extracted: (i) the UPOS tag and lemma of the matrix predicate; (ii) the UPOS tag and lemma of the embedded complement; (iii) the syntactic role of the matrix predicate (e.g., `root`, `parataxis`, `conj`); and (iv) the frequency of structural and lexical patterns (e.g., `VERB`→`xcomp(ADJ)`, *putari*→`xcomp(`*mendari*`)`).

The object-control predicate inventory comprises 17 lemmas and was derived from a frequency analysis of the 188 sentences containing `xcomp` relations. Predicates were then manually classified if they satisfied two criteria: (i) they occurred in at least one treebank instance with both a matrix object and an XCOMP complement, and (ii) the matrix object was interpreted as the implicit subject of the embedded predicate.

4.2. Controller resolution heuristics

The annotation script converts the frequency patterns identified in the XCOMP context analysis into a rule-based procedure guided by two main principles. First, **controller assignment is lexically conditioned**: a manually curated inventory (`OBJ_CTRL_VERBS`) identifies predicates that license object control, while all remaining predicates are treated as subject-control predicates, following the assumption that control is a lexical property of the matrix predicate [Bresnan et al. 2015, Kroeger 2004]. Second, **argument sharing is resolved either locally or through subject propagation**. When no overt controller is found in the local clause, the algorithm traverses `parataxis` and `conj` relations to identify controllers in higher structures, capturing recurrent pro-drop and clause-chaining patterns in Nheengatu. Table 1 summarizes the heuristic rules implemented in `xcomp_enhancer.py`.

Successful control resolution appends a dual-edge structure (`CTRL_ID:nsubj|HEAD_ID:xcomp`) to the DEPS field while preserving the original basic dependency tree. Cases that remain unresolved or already contain overt embedded subjects are automatically routed to a diagnostics log (`_skipped.tsv`) for manual review.

4.3. Manual Review

Following automatic enhancement, all annotations were manually validated using diagnostic logs for modified and unresolved cases. Manual review focused primarily on: (i) false negatives involving non-canonical object-control structures; (ii) ambiguous controllers in coordinated matrix predicates; (iii) complex `parataxis` chains; and (iv) pro-drop constructions with phonologically null controllers.

[3] All data and scripts used in this work are available at: `https://github.com/dominickmaia/EUDforNheengatu`.

Rule	Stage	Condition	Result
R1	Phase 1	Head lemma ∈ OBJ_CTRL_VERBS and overt obj/iobj present	Assign object as controller
R2	Phase 1	Overt nsubj of head (if R1 does not apply)	Assign subject as controller
R3	Phase 2	Dominating node (via parataxis/conj) has overt nsubj compatible with agrees()	Propagate subject as controller
R4	Phase 2	advcl/conj sibling of dominating node has nsubj compatible with agrees()	Propagate subject as controller
R5	Phase 2	A previously annotated xcomp is reached; controller inherited if compatible with agrees()	Propagate subject as controller
R6	Climbing	Current node is advcl with an *arama* mark dependent	Permit ascent
R7	Climbing	Current node deprel ∈ {ccomp, acl, acl:relcl}	Block ascent (clausal barrier)
R8	Pre-check	xcomp head already has an overt nsubj dependent	Skip (subject already present)
R9	Pre-check	No controller found after R1–R5	Skip (pro-drop)

Table 1. Heuristic rule mapping in `xcomp_enhancer.py`.

5. Results

The pipeline achieved 96.7% precision. Most errors involved incorrect controller propagation from higher clauses.

Type	Scope	Correct	Incorrect	Total	Precision
Subject control	Local	70	1	71	98.6%
Subject control	Propagated	14	3	17	82.4%
Object control	Local	33	0	33	100.0%
Total		**117**	**4**	**121**	**96.7%**

Table 2. Automatic annotation precision by control type and structural scope.

Among the 197 xcomp tokens, 71 (36.0%) involve subject control, 33 (16.8%) object control, and 17 (8.6%) long-distance propagation. The remaining 76 cases (38.6%) do not receive an enhanced nsubj relation because the controller is omitted (pro-drop) or requires discourse-level coreference beyond the scope of the current pipeline.

5.1. Error Analysis

The first error involves incorrect controller propagation across a parataxis chain. In sentence (3), the algorithm assigned *apigá* (‘man’) as the controller of the adjectival predicate *sasiára* (‘sad’) because it selected a compatible third-person argument while traversing the higher clause structure. However, this construction is better analyzed as object control: after the event *unupá aé* (‘hit her’), the object pronoun *aé* (‘her’) functions as the understood subject of the following predicates *uyaxiú* (‘cry’) and *upitá sasiára* (‘remain sad’). The script treats object-control separately through a manually curated inventory of object-control predicates, but this predicate was not included in that list, and the correct controller was not identified during propagation because both candidate arguments shared third-person features.

(3) *Ape , paá , nhaã apigá i piaíwa , u-nupá aé , yawé arã , paá , u-yaxiú*
then , RPRT , that man 3SG angry , 3SG-hit.IND her , so of , RPRT , 3SG-cry.IND
, u-pitá sasiára .
, 3SG-remain.IND sad .
'The man got angry and hit her, she cried and was sad.' [Casasnovas 2006, p. 89]

The second error involves incorrect controller propagation in a pro-drop configuration. In sentence (4), the algorithm incorrectly assigned *yakaré* ('alligator') as the controller of the adjectival predicate *pitunawasú* ('dark/darkness') because it selected the nearest structurally compatible third-person argument while traversing the `parataxis` chain. In this case, however, the subject of *upitá pitunawasú* ('get dark') is phonologically null, and no enhanced `nsubj` relation should be instantiated in the EUD layer.

(4) *Yawé waá , ta u-maramunha pukusá , yakaré u-suú tatá , u-puri i*
so REL , 3PL 3SG-fight.IND while , alligator 3SG-bite.IND fire , 3SG-jump.IND 3SG
irumu paraná kití , ape u-pitá pitunawasú yandé arã iké ara upé , umbaá
with river to , then 3SG-remain.IND darkness 1PL for here world in , NEG
mayé ya-maã arã .
how 1PL-see.IND PURP .
'So, while they were fighting, the alligator bit the fire and jumped into the river with it, then it got dark for us here in the world, we couldn't see.' [Avila 2021, p. 258]

The third error reveals a gap in the object-control rule. In (5), the algorithm propagated *aintá* ('they', token 5), the subject of the unrelated adjunct clause, as the controller of the `xcomp` predicate *itá* ('stone'). The correct one, however, is *aintá* ('them', token 11), the object of *remutatatinga* ('smoke'), the verb that governs the adjunct clause headed by *uyeréu* ('turn into'). Because this object is not a direct dependent of *uyeréu* but of the clause governing it, R1 does not recognize it as a local controller and fails to resolve it in Phase 1. Addressing it requires extending R1 to recognize objects of a governing clause as local controllers in adjunct clauses.

(5) *Kuíri , ti arama aintá u-musarái ne resé , re-mutatatinga aintá sikantá -pe*
now , NEG PURP 3PL 3SG-mock.IND 2SG of , 2SG-smoke.IMP 3PL resin -LOC
u-yeréu arama itá .
3SG-turn.IND PURP stone .
'Now, lest they mock you, smoke them with resin to turn them to stone.' [de Amorim 1928, p. 294]

The fourth error differs from the previous cases because it originated from an inconsistency in the original treebank annotation rather than from the enhancement procedure itself. In (6), the noun *Kurasí* ('sun') was incorrectly annotated as the `nsubj` of the parataxis predicate *asú* ('I go'), instead of being linked to the root predicate *iwaté* ('high'). As a result, the algorithm propagated *Kurasí* as the controller of the `xcomp` predicate *apurakí* ('work'). Since the head predicate is morphologically marked for first person (*a-*), a stricter person-agreement check would have blocked the incorrect propagation. The error nevertheless proved useful for identifying and revising the original annotation.

(6) *Kurasí iwaté ã , a-sú -putari a-purakí .*
sun high PFV , 1SG-go.IND -want 1SG-work.IND .
'The sun is already up, I want to go to work.' [Avila 2021, p. 350]

6. Conclusions and Future Work

The proposed pipeline achieved 96.7% precision, showing that EUD is feasible for a low-resource Indigenous language without previous EUD resources or specialized parsing tools. The pipeline also functioned as a diagnostic tool for treebank revision, identifying some pre-existing annotation inconsistencies that might otherwise have remained unnoticed. These included cases where `xcomp` dependents incorrectly contained overt `nsubj` dependents, as well as errors involving the attachment of `xcomp` relations to the wrong predicate head, improving the internal consistency of the treebank itself.

This work shows that EUD annotation for endangered and low-resource languages can be developed through rule-based methods combining corpus-based heuristics and manual validation. Future work includes enriching the JSON glossary used by the Yauti tool with valency information, integrating the enhancer into the Yauti automatic UD annotation tool [de Alencar 2023], and extending the approach to additional EUD phenomena beyond `xcomp` relations. The resulting guidelines also provide a starting point for future inter-annotator agreement experiments and for expanding EUD resources for Indigenous languages within the UD ecosystem.

We also plan to investigate the adaptation of the rule-based graph rewriting system of [Guillaume and Perrier 2021] to Nheengatu by using language-specific lexical information for control and raising predicates. This will allow us to compare this approach with the Python-based pipeline presented here in terms of annotation quality, computational complexity, processing speed, and ease of integration with the existing annotation tool, Yauti.

Limitations

The reported frequencies of object-control predicates reflect only the current version of the treebank, and as new sentences are annotated, both predicate frequencies and the inventory may change.

Furthermore, the pipeline was developed and applied by a single annotator experienced in UD annotation for Nheengatu, who also manually reviewed the 121 automatically generated EUD annotations and analyzed the annotation errors. The verb classifications were reviewed by a second experienced annotator. However, the annotation guidelines were not validated through an inter-annotator agreement experiment, and their consistency across independent annotators remains to be assessed.

Ethics Statement

All data used in this study come from the UD_Nheengatu-CompLin treebank, distributed as part of Universal Dependencies v2.17 [Zeman et al. 2025]. No new data involving human participants was collected. We recognize that computational research on Indigenous languages involves responsibilities toward the communities connected to these languages, and we support efforts to develop accessible and reusable NLP resources beyond academic contexts.

Generative AI disclosure

We used generative AI tools, including Claude and Gemini, to help write the Python scripts, assist with LaTeX editing and formatting, correct grammatical errors, and polish the overall clarity of the text. All linguistic analysis, annotation decisions, rule creation, error analysis, and final code verification were carried out by the authors.

Acknowledgements

We thank the anonymous reviewers for their insightful feedback and suggestions, which helped improve the quality of this paper.

References

Abeillé, A. (2024). Control and raising. In Müller, S., Abeillé, A., Borsley, R. D., and Koenig, J.-P., editors, *Head-Driven Phrase Structure Grammar: The Handbook*, number 9 in Empirically Oriented Theoretical Morphology and Syntax, pages 519–570. Language Science Press, Berlin, 2 edition.

Alencar, L. F. d., Silva, H. L. B., Gurgel, J. L., and Alexandre, D. M. (2025). *Per aspera ad astra*: Improving the automatic evaluation of a Universal Dependencies treebank for a low-resourced language. *Revista de Estudos da Linguagem*. In press.

Avila, M. T. (2021). *Proposta de dicionário nheengatu-português*. PhD thesis, Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo.

Bresnan, J., Asudeh, A., Toivonen, I., and Wechsler, S. (2015). *Lexical-functional syntax*, volume 16. John Wiley & Sons.

Casasnovas, A. (2006). *Noções de língua geral ou nheengatú: gramática, lendas e vocabulário*. Editora da Universidade Federal do Amazonas; Faculdade Salesiana Dom Bosco, Manaus, 2 edition.

Church, K. and Liberman, M. (2021). The future of computational linguistics: On beyond alchemy. *Frontiers in Artificial Intelligence*, 4:1–18.

de Alencar, L. F. (2023). Yauti: A tool for morphosyntactic analysis of Nheengatu within the Universal Dependencies framework. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 135–145, Porto Alegre, RS, Brasil. SBC.

de Alencar, L. F. (2024). A Universal Dependencies treebank for Nheengatu. In Gamallo, P., Claro, D., Teixeira, A. J. S., Real, L., García, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, volume 2, pages 37–54, Santiago de Compostela, Galicia, Spain. Association for Computational Linguistics.

de Alencar, L. F. (2025). Enhancing a nheengatu morphosyntactic analyzer for word formation and non-standard language. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 13–28, Porto Alegre, RS, Brasil. SBC.

de Alencar, L. F. and Rademaker, A. (2022). Modelação da valência verbal numa gramática computacional do português no formalismo hpsg. *Domínios de Lingu@gem*, 16(4):1339–1400.

de Amorim, A. B. (1928). Lendas em Nheêngatu e em Portuguez. *Revista do Instituto Historico e Geographico Brasileiro*, 154(100):9–475. Tomo 100, vol. 154 (2º de 1926).

de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C., McDonald, R., Nivre, J., Petrov, S., Pyysalo, S., Schuster, S., Silveira, N., Tsarfaty, R., Tyers, F., and Zeman, D. (2025). xcomp: open clausal complement. `https://universaldependencies.org/u/dep/xcomp.html`.

Di Felippo, A., Postali, C., Ceregatto, G., Gazana, L., Silva, E., Roman, N., and Pardo, T. (2021). Descrião preliminar do corpus DANTEStocks: Diretrizes de segmentação para anotação segundo Universal Dependencies. In Ruiz, E. E. S. and Torrent, T. T., editors, *Proceedings of the 13th Brazilian Symposium in Information and Human Language Technology*, pages 335–343, Porto Alegre, Brazil. Association for Computational Linguistics.

Duran, M., Lopes, L., das Graças Nunes, M., and Pardo, T. (2023). The dawn of the porttinari multigenre treebank: Introducing its journalistic portion. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 115–124, Porto Alegre, RS, Brasil. SBC.

Guillaume, B. and Perrier, G. (2021). Graph rewriting for enhanced Universal Dependencies. In Oepen, S., Sagae, K., Tsarfaty, R., Bouma, G., Seddah, D., and Zeman, D., editors, *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 175–183, Online. Association for Computational Linguistics.

Hartt, C. F. (1938). Notas sobre a língua geral, ou tupí moderno do Amazonas. *Anais da Biblioteca Nacional do Rio de Janeiro*, LI:305–390. [1929].

Kroeger, P. (2004). *Analyzing Syntax: A Lexical-Functional Approach*. Cambridge University Press, Cambridge.

Navarro, E. d. A. (2012). O último refúgio da língua geral no Brasil. *Estudos Avançados*, 26(76):245–254.

Nivre, J. et al. (2016). Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Pagano, A. S., Duran, M. S., and Pardo, T. A. S. (2023). Enhanced dependencies para o português brasileiro. In Pardo, T. A. S., Duran, M. S., and Lopes, L., editors, *Proceedings of the 2nd Edition of the Universal Dependencies Brazilian Festival*, pages 461–470, Belo Horizonte, Brazil. Association for Computational Linguistics.

Pugh, R., Huerta Mendez, M., Sasaki, M., and Tyers, F. (2022). Universal Dependencies for western sierra Puebla Nahuatl. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5011–5020, Marseille, France. European Language Resources Association.

Pugh, R. and Tyers, F. (2024). A Universal Dependencies treebank for Highland Puebla Nahuatl. In Duh, K., Gomez, H., and Bethard, S., editors, *Proceedings of the 2024*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 1393–1403, Mexico City, Mexico. Association for Computational Linguistics.

Rodrigues, J. B. (1890). *Poranduba amazonense ou kochiyma-uara porandub, 1872-1887*. Typ. de G. Leuzinger & Filhos, Rio de Janeiro.

Sanches Duran, M., de Souza, E. A., das Graças Volpe Nunes, M., Pagano, A. S., and Pardo, T. A. S. (2025). Extending the enhanced Universal Dependencies – addressing subjects in pro-drop languages. In Bouma, G. and Çöltekin, Ç., editors, *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 143–152, Ljubljana, Slovenia. Association for Computational Linguistics.

Schuster, S. and Manning, C. D. (2016). Enhanced English Universal Dependencies: An improved representation for natural language understanding tasks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2371–2378, Portorož, Slovenia. European Language Resources Association (ELRA).

Souza, E., Duran, M., Nunes, M. d. G. V., Sampaio, G., Belasco, G., and Pardo, T. (2024). Automatic annotation of enhanced Universal Dependencies for Brazilian Portuguese. In Claro, D. B. and Pagano, A., editors, *Proceedings of the 15th Brazilian Symposium in Information and Human Language Technology*, pages 22–31, Belém do Pará, Brazil. Association for Computational Linguistics.

Thomas, G. (2019). Universal Dependencies for Mbyá Guaraní. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 70–77, Paris, France. Association for Computational Linguistics.

Universal Dependencies (2026). Enhanced Universal Dependencies syntax overview. `https://universaldependencies.org/u/overview/enhanced-syntax.html`. Accessed: 2026-05-01.

Zeman, D. et al. (2025). Universal dependencies 2.17. Accessed: 2026-05-14.

Leakage-Aware Evaluation of Brazilian Clinical Notes for ICU Mortality Prediction: A Study on the BRATECA Dataset

Felipe André Bach Alves[1], **Heloísa Oss Boll**[2], **Mariana Recamonde-Mendoza**[1,3]

[1]Institute of Informatics, Universidade Federal do Rio Grande do Sul (UFRGS), Brazil

[2]Dept. of Medical Informatics, Amsterdam UMC/University of Amsterdam, The Netherlands

[3]Bioinformatics Core, Hospital de Clínicas de Porto Alegre (HCPA), Brazil

{mrmendoza, felipe.alves}@inf.ufrgs.br, h.ossboll@amsterdamumc.nl

Abstract. *Early prediction of ICU mortality can support clinical decisions, but retrospective notes can contain future-event cues that cause leakage in clinical NLP. We evaluate leakage-aware modeling of Brazilian Portuguese ICU notes for mortality prediction using BRATECA. We compare 24-hour restriction, regex and LLM leakage auditing, neural and TF-IDF representations, and unrestricted-note baselines. In the 24-hour setting, models achieved AUROC 0.78-0.83; unrestricted notes reached around 0.95, indicating outcome-related information in late documentation. Regex found leakage signals in 21.13% of the notes, and the LLM audit identified additional semantic cases. Results highlight the need for temporal control and leakage auditing in clinical NLP.*

1. Introduction

Predicting clinical outcomes from *electronic health records* (EHRs) is a central task in medical *machine learning* (ML), particularly in high-risk settings such as *intensive care units* (ICUs). Early prediction of in-hospital mortality may support clinical decision-making, resource allocation, and risk stratification [Choi et al. 2022]. EHR datasets combine structured information, such as demographics, examinations, and prescriptions, with unstructured clinical narratives that capture clinical reasoning and patient trajectories. Recent studies have emphasized the importance of leveraging multimodal EHR data to improve predictive modeling and representation learning in healthcare [Boll et al. 2024].

Clinical notes are especially valuable because they include contextual information often absent from structured fields [Seinen et al. 2025]. However, they also introduce methodological risks: notes may be written retrospectively, copied forward, amended after an event, or produced near discharge or death, so they may contain information unavailable at the intended prediction time. This is a form of data leakage, that is, the model receiving information unavailable at the intended prediction time. For instance, when predicting mortality from documentation available within the first 24 hours of admission, a note written days later stating that the patient died or was discharged reveals the target and must be excluded as input. Used without temporal control, such notes let models exploit outcome-related cues instead of early clinical signals, yielding overly optimistic estimates and poor prospective validity [Chiavegatto Filho et al. 2021].

This risk is further intensified in ICU environments, where patients exhibit rapidly evolving clinical trajectories, dense monitoring, frequent interventions, and extensive documentation histories. Despite growing interest in clinical text modeling, leakage-aware

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 14–25
October 19-22, 2026 ©2026 Association for Computational Linguistics

analyses of ICU narratives remain limited, particularly beyond English-language benchmark datasets such as MIMIC [van Aken et al. 2021]. In this context, clinical narratives in Brazilian Portuguese remain underexplored, and their documentation practices may differ substantially from those represented in widely used international benchmarks.

This paper investigates data leakage in ICU clinical notes for in-hospital mortality prediction using the Brazilian clinical data collection BRATECA [Dias and Ulbrich 2022, Consoli et al. 2022]. Because the dataset does not currently include metadata about the note types, robust leakage auditing is essential before such notes can be reliably incorporated into prediction pipelines. Therefore, our central *natural language processing* (NLP) questions are: 1) To what extent does the predictive performance of models using these clinical notes for ICU mortality prediction indicate real, early clinical signal rather than retrospective outcome leakage? and 2) How can temporal restriction and semantic auditing be combined to help quantify this gap in these clinical narratives?

Rather than proposing a new predictive architecture, this study makes a methodological and empirical contribution to clinical NLP. Specifically, we: (i) quantify how temporal availability of a Brazilian Portuguese clinical note dataset changes estimated mortality-prediction performance; (ii) compare lexical, dense, Portuguese-language, and translated-English representations under controlled prediction windows; and (iii) combine rule-based and LLM auditing to characterize explicit and semantic outcome cues. Our results show that conclusions about predictive quality are highly sensitive to how temporal cuts and eligible documentation are defined, offering concrete guidance for avoiding leakage-driven overestimation on BRATECA or other clinical notes datasets.

2. Related Work

Predictive modeling using ML and EHR data has been extensively studied for ICU risk assessment, particularly for mortality prediction. Most existing models rely primarily on structured variables such as vital signs, laboratory results, prescriptions, and severity scores derived from datasets such as MIMIC [Johnson et al. 2016], frequently achieving performance comparable to or exceeding traditional clinical scoring systems such as APACHE, SAPS, and SOFA [Knaus et al. 1985, Le Gall et al. 1993, Vincent et al. 1996, Olang et al. 2025]. Recent work has also shown the value of combining structured variables with clinical narratives, using multimodal EHR representations for risk prediction [Boll et al. 2024, Garriga et al. 2023].

Clinical NLP has enabled systematic use of free-text documentation through lexical and representation-learning methods. Lexical baselines such as *term frequency–inverse document frequency* (TF-IDF) remain important in clinical NLP because they provide strong, interpretable comparisons and can expose reliance on explicit outcome words [Tavabi et al. 2024]. More recent studies rely on contextual representations from transformer-based models such as BioBERT and ClinicalBERT [Alsentzer et al. 2019, Lee et al. 2020]. These approaches can capture semantic patterns in clinical notes and have been applied to mortality prediction and related ICU tasks [Vagliano et al. 2023]. More recently, general and domain-adapted decoder models, including ChatGPT, Portuguese *large language models* (LLMs) such as Sabiá, and medical LLMs such as Meditron, have expanded the range of NLP methods available for clinical and biomedical tasks [OpenAI 2023, Nogueira et al. 2023, Chen et al. 2023].

However, data leakage remains a persistent challenge in health ML. Prior work has discussed leakage as a source of biased performance estimates and poor reproducibility [Chiavegatto Filho et al. 2021]. In clinical narratives, leakage may arise when documentation includes information generated after the intended prediction time. Although methodological frameworks for the development of leakage-aware models have been proposed [Davis et al. 2023], the specific problem of temporal leakage in non-English ICU notes remains not sufficiently characterized. Here, we address this gap by exploring multiple strategies for leakage-aware modeling of Brazilian ICU narratives, including an LLM component for semantic leakage auditing in Portuguese.

3. Methodology

3.1. Dataset, ICU Cohort, and Prediction Task

We used BRATECA, a Brazilian retrospective EHR dataset containing relational tables with admissions, clinical notes, exams, prescriptions, and prescription items [Dias and Ulbrich 2022, Goldberger et al. 2000]. Table 1 summarizes the source tables used in this study. The target prediction task was in-hospital mortality. Figure 1 summarizes the cohort construction, clinical-note preprocessing, and three experimental branches used in this study.

The primary source of unstructured text was *B1_ClinicalNote*. This table contains heterogeneous Brazilian Portuguese free-text documentation produced during hospitalization, including notes consistent with admission documentation, clinical progress/evolution, nursing and medical updates, procedure-related records, discharge-oriented documentation, and administrative or care-transition records. These narratives vary in length, style, abbreviation density, and temporal proximity to clinical events, which makes them useful for clinical NLP while also increasing the risk of temporal and outcome-related leakage.

Table 1. BRATECA source tables used in this study.

Table	Cols	Rows	Description
B1_Admission	11	72,085	Encounter-level admission data, including patient and hospital identifiers and discharge information.
B1_ClinicalNote	6	2,855,818	Free-text clinical notes recorded during hospitalization.
B1_Exam	7	2,372,088	Exams per admission and their results.
B1_Prescription	30	518,420	Prescription headers, timestamps, and ICU-related flags.
B1_PrescriptionItem	30	5,737,707	Prescription items linked to prescription headers.

The ICU cohort was identified using the ICU flag in the prescription table: an admission was considered an ICU encounter when it appeared in *B1_Prescription* with at least one record where *IC* was true. This definition yielded 4,779 ICU admissions. Outcome labels were extracted from discharge disposition values and binarized as death versus discharge alive. Dispositions outside the modeling scope, such as transfers and non-standard discharges, were excluded. The modeled death-versus-discharge cohort contained 4,723 ICU admissions, including 1,098 deaths and 3,625 discharges alive.

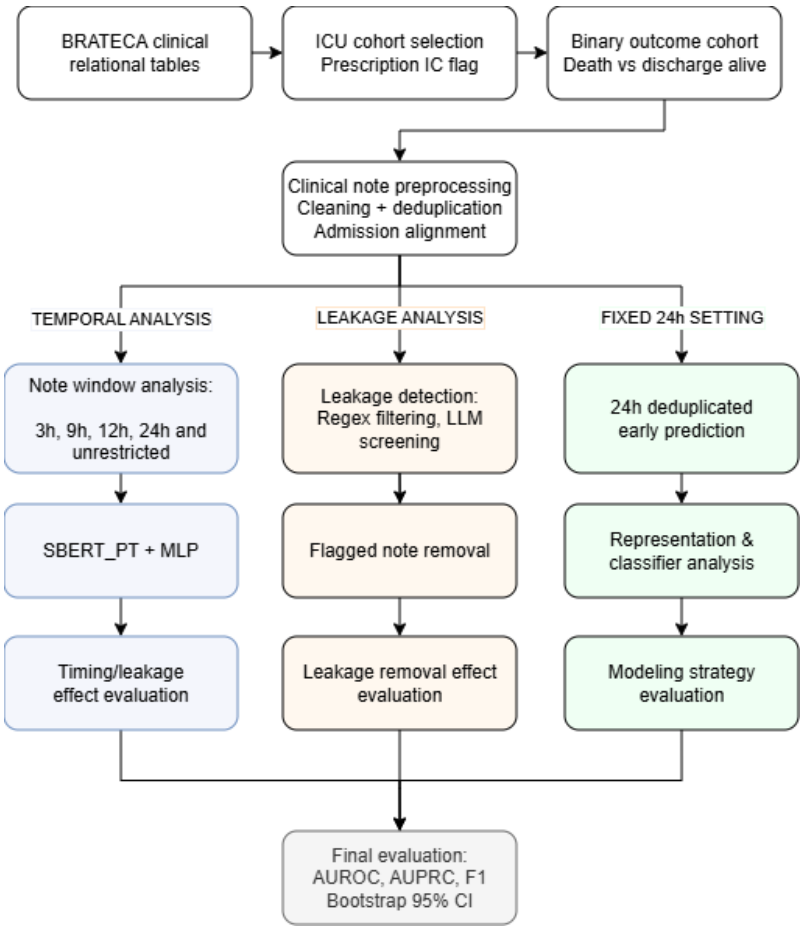


Figure 1. Overview of cohort construction and the temporal, leakage-audit, and fixed 24-hour evaluation branches.

In the modeled death-versus-discharge ICU cohort, the recorded time from admission to discharge was 18.13 $\pm$ 17.50 days (median 12.72; *interquartile range* (IQR) 6.98-23.12). Among the modeled admissions with clinical notes, the last available clinical note occurred 19.06 $\pm$ 17.99 days after admission (median 13.74; IQR 7.47-24.45), showing that unrestricted note sets frequently include documentation from late hospitalization stages. Because the main text-only experiments required at least one eligible clinical note within the 24-hour window after deduplication and embedding alignment, the aligned 24-hour matrix contained 4,668 admissions and 1,078 deaths.

Table 2 summarizes the main raw, derived, and generated fields used in the experiments. Identifiers and outcome labels were used for linkage, splitting, cohort definition, or evaluation only; they were not used as text-model predictors and were not submitted to the LLM audit.

3.2. Leakage Mitigation

We evaluated a leakage-aware preprocessing strategy with four components. First, temporal windowing restricted notes to clinically plausible prediction horizons, especially the first 24 hours after hospital admission among ICU-flagged admissions. Second, rule-based regular expressions flagged notes containing explicit or proxy outcome-related expressions, including terms related to death, discharge, transfers, failed resuscitation, ter-

Table 2. Raw, derived, and generated fields used in the experimental pipeline.

Field or group	Type	Role in the study
Admission_ID, *Patient_ID*, *Hospital_ID*	Raw identifiers	Used for relational linkage, admission-level aggregation, and split construction. They were not used as predictive features and were not submitted to the LLM.
IC	Raw structured flag	Used only to define ICU admissions from the prescription table.
Admission_Date, *Note_Date*, hours since admission	Raw/derived time fields	Used to compute temporal eligibility windows such as 3h, 9h, 12h, and 24h. These time fields were not included as text-model predictors.
Note_Text_Clean	Derived text	Cleaned Brazilian Portuguese clinical note text used to build embeddings, TF-IDF features, deduplication keys, and LLM audit inputs.
target_obito	Derived outcome	Binary in-hospital mortality label derived from discharge disposition values. It was used for training/evaluation but was not submitted to the LLM audit.
Regex leakage flag	Generated audit field	Rule-based indicator of explicit or proxy outcome-related expressions, used for leakage prevalence estimation and filtering sensitivity analyses.
Embedding vectors	Generated features	Dense note-level representations generated by pretrained encoders and aggregated to the admission level, mainly through mean pooling after deduplication.
LLM audit fields	Generated audit fields	Structured fields such as leakage presence, leakage type, evidence, explanation, and confidence. They were used only for exploratory leakage auditing and sensitivity analysis, not as predictors in the mortality model.

minal documentation, and late addenda. Third, the last note of each ICU admission was excluded in selected configurations because final notes often summarize the trajectory or outcome. Fourth, duplicated clinical notes within the same admission were removed using cleaned note text, reducing the influence of repeated or copied documentation.

The experiments used the preprocessed `Note_Text_Clean` field. The upstream cleaning procedure removed markup artifacts and textual noise and applied character normalization. The same cleaned field was used for deduplication, representation learning, TF-IDF vectorization, and LLM audit inputs. Table 3 illustrates the main lexical categories targeted by the regex audit. The examples are synthetic Brazilian Portuguese clinical-text cues derived from the rule categories, not excerpts from BRATECA.

Table 3. Synthetic examples of lexical leakage cues targeted by the regex audit.

Category	Synthetic lexical cue	Potential leakage mechanism
Explicit death	"evoluiu a obito"; "obito constatado"	Directly reveals the mortality outcome used as the prediction label.
Discharge	"alta hospitalar"; "recebeu alta"	Reveals survival/discharge information that should not be available at early prediction time.
Transfer or post-discharge trajectory	"transferido para enfermaria"; "home care"	Indicates later care trajectory or disposition after ICU documentation.
Failed resuscitation or vital-status proxy	"PCR sem retorno"; "sem sinais vitais"	Provides a strong proxy for death even when the outcome label is not explicitly named.
Terminal context	"cuidados paliativos"; "conforto exclusivo"	May reveal end-of-life trajectory or late-stage clinical decisions.
Administrative closure	"caso encerrado"; "desfecho"	Suggests documentation created after the hospitalization outcome was known.

Regex screening was designed as a baseline audit of lexical leakage indicators, targeting explicit outcome-related expressions and predefined patterns. To complement this approach, we also performed an exploratory LLM-based semantic audit aimed at identifying potentially outcome-revealing content that may not be captured by surface lexical cues alone. This evaluation uses the Azure AI Foundry OpenAI-compatible Responses API with the model `gpt-5.4-mini`, version 2026-03-17, with temperature = 0.1 and structured JSON Schema outputs. This deployment was selected for scalable, monitored batch processing with structured JSON outputs and latency/cost compatible

with full-corpus screening. Only cleaned clinical-note text was submitted, and under the enterprise/API configuration used in this study, prompts and outputs are not reused to train foundation models.

3.3. Embeddings and Aggregation

We evaluated multiple pretrained embedding configurations across language and domain settings. The multilingual sentence-level baseline used `sentence-transformers/paraphrase-multilingual-mpnet-base-v2` (SBERT), which produces 768-dimensional note representations. Additional configurations used BERTimbau for Brazilian Portuguese, Portuguese and English BioBERT variants, and ClinicalBERT for English clinical text.

For English-language configurations, Brazilian Portuguese notes were translated using `facebook/nllb-200-distilled-600M` [Costa-jussà et al. 2022]. Translation was performed in batches and persisted as Parquet shards, which were subsequently concatenated while preserving record order and alignment with the original note identifiers. This alignment allowed the same temporal filters, labels, and admission-level aggregation procedure to be applied to the original and translated notes. After within-admission deduplication and temporal filtering, note-level embeddings were aggregated into a single admission-level representation using mean pooling.

3.4. Experimental Protocol

We evaluated in-hospital mortality prediction at the admission level. The main tables use a stratified 70/20/10 admission-level train/validation/test split with a fixed random seed. The main neural classifier was a multilayer perceptron (MLP) trained on admission-level embeddings. The MLP consisted of two hidden layers with 256 and 64 units, ReLU activations, dropout of 0.30 and 0.20, and a single output logit. Models were trained with AdamW, learning rate 10^{-3}, weight decay 10^{-4}, batch size 128, and early stopping based on validation AUROC. Class imbalance was handled using cost-sensitive learning.

For comparison, we evaluated logistic regression, linear *support vector machine* (SVM), *Gaussian naive Bayes* (NB), and nearest centroid using the same SBERT_PT 24-hour admission-level matrix as the MLP. TF-IDF word 1-2-gram and character 3-5-gram baselines used the same eligible admissions and split, but their own sparse representations, with class-balanced logistic regression. Thresholds were selected using the validation set. We report AUROC, AUPRC, F1, precision, recall, and balanced accuracy; 95% confidence intervals were estimated from 2000 bootstrap resamples of test admissions with replacement. The five-seed LLM sensitivity analysis is reported separately.

4. Results and Discussion

4.1. Prevalence of Regex-Detected Leakage Signals

Regex-based leakage screening identified 261,395 notes with explicit or proxy outcome-related expressions among 1,236,991 evaluated notes, corresponding to 21.13% of the note corpus (Table 4). This confirms that leakage indicators are frequent in clinical narratives and motivates temporal and linguistic auditing.

Table 4. Prevalence of leakage signals identified by regex-based auditing.

Leakage detected	Notes	Share
No	975,596	78.87%
Yes	261,395	21.13%
Total	1,236,991	100.00%

4.2. Comparison of Embedding Representations for Mortality Prediction

Table 5 shows the 24-hour embedding comparison for the in-hospital mortality prediction task using an MLP classifier. Performance differs across embedding models, indicating that the choice of clinical note encoder impacts mortality prediction. SBERT-based representations obtained the highest observed AUROC values, but SBERT_EN and SBERT_PT were nearly tied and their bootstrap intervals overlap. Therefore, the small AUROC difference should not be interpreted as robust superiority. SBERT_EN achieved higher observed AUPRC in this configuration, whereas SBERT_PT remains methodologically attractive because it operates directly on the original Brazilian Portuguese notes without machine translation, avoiding translation costs and possible semantic shifts.

Table 5. Performance of different embedding methods for ICU mortality prediction under the 24-hour setting using an MLP classifier.

Model	AUROC (95% CI)	AUPRC (95% CI)	F1
SBERT_EN	0.8033 [0.760, 0.848]	0.5927 [0.497, 0.690]	0.5205
SBERT_PT	0.8025 [0.755, 0.849]	0.5676 [0.475, 0.664]	0.5190
BERTimbau_PT	0.7810 [0.733, 0.834]	0.4978 [0.402, 0.598]	0.5432
BioBERT_PT	0.7653 [0.711, 0.817]	0.4487 [0.363, 0.547]	0.5179
BioBERT_EN	0.7506 [0.707, 0.802]	0.4316 [0.350, 0.527]	0.4910
ClinicalBERT_EN	0.7482 [0.702, 0.803]	0.4719 [0.381, 0.569]	0.4873

4.3. Performance of Lexical TF-IDF Baselines

Because lexical cues are central to leakage analysis in clinical notes, we evaluated TF-IDF logistic-regression (LR) baselines on the same 24-hour admission-level split used for the embedding experiments. Unlike previous neural representations, which encode each note as a dense semantic vector before admission-level pooling, TF-IDF represents admissions through sparse word- and character-level n-gram patterns. Table 6 shows that both word- and character-level TF-IDF baselines were competitive with, and in this holdout exceeded, the neural embedding probes reported in Table 5. This result strengthens the NLP interpretation of the task: explicit lexical patterns in Brazilian Portuguese clinical notes retain considerable predictive information even under early temporal restriction.

Table 6. Performance of lexical TF-IDF baselines under the aligned 24-hour setup.

Model	AUROC (95% CI)	AUPRC (95% CI)	F1
TF-IDF word 1-2g + LR	0.8293 [0.788, 0.873]	0.6049 [0.514, 0.702]	0.5674
TF-IDF char 3-5g + LR	0.8310 [0.790, 0.873]	0.6221 [0.533, 0.715]	0.5783

4.4. Analysis of Temporal-Cutoff Sensitivity

Figure 2 reports the sensitivity of text-only SBERT_PT performance to note availability over time. Performance varied non-monotonically across the early prediction windows, with overlapping confidence intervals and differences in the amount of eligible documentation. In contrast, the unrestricted setting produced a substantially larger increase in both AUROC and AUPRC. This does not imply that early notes lack useful signal; rather, it shows that unrestricted retrospective text can inflate performance estimates when temporal availability is not controlled.

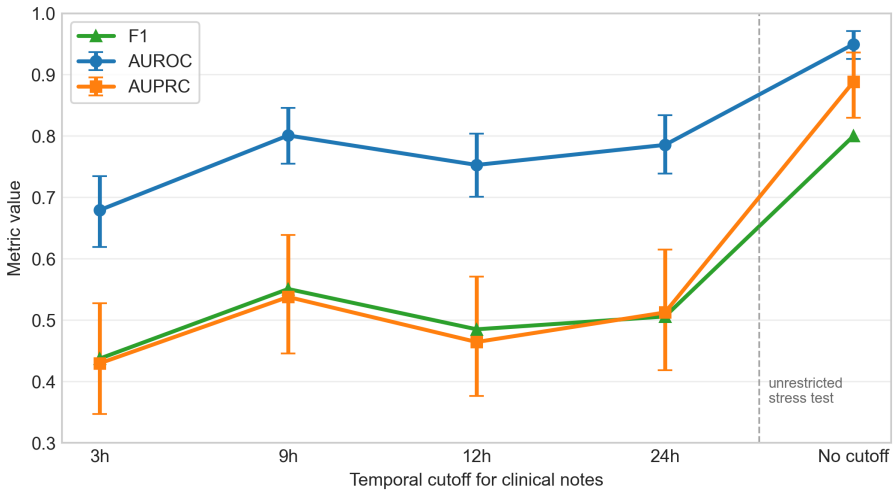


Figure 2. Text-only temporal cutoff sensitivity for SBERT_PT using mean-pooled, deduplicated note embeddings, with 95% CI (AUROC and AUPRC).

The "No cutoff" setting in Figure 2 and the "Unrestricted notes - No regex cleaning" setting in Table 8 are both unrestricted-text upper bounds, but they are not identical. Figure 2 uses the temporal-sensitivity feature lineage derived from the modeled ICU cohort of 4,723 admissions and 1,098 deaths, where as the regex-ablation pipeline reconstructs the unrestricted corpus from the regex lineage with 4,722 eligible admissions and 1,098 deaths. Therefore, the small numerical difference between AUROC 0.9496 and 0.9458 reflects slightly different eligibility/source lineage and split composition; however, the intended interpretation is the same: unrestricted retrospective notes yield substantially inflated performance.

4.5. Comparison of Traditional Classifiers Under the 24-Hour Setting

To compare classifier families under a fixed representation, we used a separately rebuilt SBERT PT 24-hour feature matrix aligned across all classifiers. To address whether neural models outperform simpler classifiers under the same representation, we compared the MLP with traditional baselines using the same SBERT_PT 24-hour mean-pooled, deduplicated admission-level matrix in the in-hospital mortality prediction task. Table 7 shows that the MLP achieved the highest observed AUROC and AUPRC, indicating better ranking performance over the shared embedding representation. Logistic regression was the closest traditional baseline, with lower AUROC/AUPRC but a similar F1, driven by higher recall and lower precision. The remaining traditional methods performed substantially worse, especially in AUPRC, suggesting a weaker ability to rank positive mortality cases

under class imbalance. Under the shared SBERT_PT representation, the MLP had the highest observed AUROC and AUPRC, while logistic regression achieved a similar F1 with higher recall and lower precision. Because the closest bootstrap intervals overlap, these results do not establish statistically robust superiority of the MLP.

Table 7. Aligned comparison between MLP and traditional classifiers using the same rebuilt SBERT_PT 24-hour text-only feature matrix.

Model	AUROC (95% CI)	AUPRC (95% CI)	F1
MLP probe	0.7855 [0.739, 0.834]	0.5125 [0.418, 0.615]	0.5059
Logistic regression	0.7527 [0.699, 0.808]	0.4806 [0.393, 0.581]	0.5112
Linear SVM	0.7113 [0.656, 0.769]	0.4128 [0.335, 0.510]	0.4944
Gaussian NB	0.6939 [0.639, 0.749]	0.3459 [0.279, 0.419]	0.4215
Nearest centroid	0.6827 [0.628, 0.738]	0.3407 [0.277, 0.414]	0.4000

4.6. Impact of Regex-Based Leakage Filtering

Table 8 compares prediction performance with and without regex-based filtering. In the 24-hour setting, filtering produced a small observed improvement, but the bootstrap intervals overlap and the result should not be interpreted as a robust gain. A plausible explanation is removal of noisy, copied, or administratively late documentation, but this remains a sensitivity finding. In the unrestricted setting, filtering reduced performance, consistent with the removal of outcome-related content from late-stage notes. These findings show that leakage is not simply the presence of outcome-correlated terms, but the use of information unavailable at prediction time. Regex filtering can mitigate explicit lexical cues, but temporal restriction addresses the root cause more directly.

Table 8. Impact of regex-based leakage filtering on SBERT_PT performance. Unrestricted notes include all eligible documentation recorded from hospital admission through discharge or death, including notes written after the 24-hour prediction cutoff. These configurations represent retrospective upper bounds rather than deployable early-prediction estimates.

Configuration	AUROC (95% CI)	AUPRC (95% CI)	F1
24h window - No regex filtering	0.7518 [0.702, 0.800]	0.4371 [0.354, 0.534]	0.5139
24h window - With regex filtering	0.7603 [0.706, 0.811]	0.4732 [0.385, 0.581]	0.5177
Unrestricted notes - No regex filtering	0.9458 [0.918, 0.966]	0.8844 [0.826, 0.929]	0.8019
Unrestricted notes - With regex filtering	0.9332 [0.905, 0.956]	0.8314 [0.762, 0.887]	0.7449

4.7. Exploratory Semantic Leakage Auditing with LLMs

In the preliminary audit, 61,852 ICU note records were evaluated and 2,430 were marked as containing potential leakage. In five-seed sensitivity analysis, LLM-based exclusion did not robustly change aggregate performance; however, among admissions affected by flagged notes, MLP AUPRC decreased from 0.8735 to 0.8321 after exclusion.

Among the flagged notes, 197 did not match any regex rule and most often contained discharge, indirect-outcome, or terminal-context cues. These cases illustrate how semantic auditing may complement predefined lexical patterns, although the annotations were not validated against a human gold standard.

5. Limitations

This study has several limitations. First, experiments used a single Brazilian ICU dataset (BRATECA), which may limit generalizability to other healthcare systems, languages, and documentation practices. Second, regex-based leakage detection captures explicit and proxy lexical cues but cannot exhaustively identify semantic or context-dependent outcome information; our preliminary LLM audit surfaced additional cases but was not validated against a human-annotated gold standard. Third, differences between Brazilian Portuguese and English representations may reflect both encoder behavior and translation artifacts from the NLLB-200 preprocessing step. Fourth, central comparisons rely on a single fixed admission-level train/validation/test split with bootstrap uncertainty rather than repeated cross-validation. Because splitting occurred at the admission level, admissions from the same patient may fall in different partitions; quantifying this overlap and testing patient-grouped splits is important future work. Sample sizes also vary slightly across experiments, since temporal cutoffs, leakage filtering, and deduplication remove different subsets of notes. Broader cross-dataset validation is needed to determine whether the leakage patterns observed here generalize beyond BRATECA.

6. Conclusion

This study evaluated distinct strategies for leakage-aware modeling of Brazilian ICU clinical notes for in-hospital mortality prediction using BRATECA. The results show that early clinical narratives contain meaningful predictive signal, with 24-hour text-based models achieving AUROC around 0.78-0.83 depending on whether neural embeddings or lexical TF-IDF representations are used. However, including unrestricted retrospective notes substantially increases text-only performance to AUROC around 0.95, which indicates the presence of outcome-related information in the documentation.

The findings support three methodological conclusions. First, strict temporal control is essential when using clinical notes for prediction. Second, regex-based filtering can identify frequent leakage indicators but should be treated as a partial mitigation strategy. Third, LLM-based semantic auditing is promising for identifying subtler forms of leakage. A qualitative inspection of LLM-only cases suggests a path for extending this research toward LLM-based mortality prediction. For example, some notes flagged only by the LLM described positive clinical and apnea tests compatible with a brain-death protocol. Although such notes may not contain the exact regex triggers for death or discharge, they semantically reveal a terminal trajectory and information unavailable at early prediction time. This qualitative inspection also motivates future work on temporally constrained LLM-based representations for mortality prediction. Overall, this work emphasizes that reliable clinical-text modeling requires not only predictive performance reporting, but also explicit auditing of what information was available at the intended prediction time.

Ethics Statement

The BRATECA dataset was accessed through PhysioNet under credentialed access and is distributed as deidentified research data. This study did not attempt to reidentify patients, clinicians, hospitals, or institutions. For the exploratory LLM audit, only cleaned clinical-note text was submitted to Azure AI Foundry; under the enterprise/API configuration used, submitted prompts and outputs are not reused for foundation-model training.

Disclosure of Generative AI Usage

Generative artificial intelligence tools were used to support language revision, manuscript organization, and the summarization of experimental findings. All scientific decisions, analyses, interpretations, and conclusions were critically reviewed by the authors and remain their sole responsibility.

References

Alsentzer, E., Murphy, J., Boag, W., et al. (2019). Publicly available clinical bert embeddings. In *Proceedings of the Clinical NLP Workshop*.

Boll, H. O., Amirahmadi, A., Ghazani, M. M., de Morais, W. O., de Freitas, E. P., Soliman, A., Etminani, F., Byttner, S., and Recamonde-Mendoza, M. (2024). Graph neural networks for clinical risk prediction based on electronic health records: A survey. *Journal of Biomedical Informatics*.

Chen, Z., Hernández Cano, A., Romanou, A., Bonnet, A., Matoba, K., Salvi, F., Pagliardini, M., et al. (2023). Meditron-70b: Scaling medical pretraining for large language models.

Chiavegatto Filho, A. D. P., Batista, A. F., and Santos, H. G. (2021). Data leakage in health outcomes prediction with machine learning. comment on "prediction of incident hypertension within the next year: Prospective study using statewide electronic health records and machine learning". *Journal of Medical Internet Research*, 23(2).

Choi, M. H., Kim, D., Choi, E. J., Jung, Y. J., Choi, Y. J., Cho, J. H., and Jeong, S. H. (2022). Mortality prediction of patients in intensive care units using machine learning algorithms based on electronic health records. *Scientific Reports*, 12(1):7180.

Consoli, B. S., dos Santos, H. D. P., Ulbrich, A. H. D. P. S., Vieira, R., and Bordini, R. H. (2022). Brateca (brazilian tertiary care dataset): a clinical information dataset for the portuguese language. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022)*, Marseille, France. European Language Resources Association (ELRA).

Costa-jussà, M. R., Cross, J., Červeňansky, F., et al. (2022). No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Davis, S. E., Matheny, M. E., Balu, S., and Sendak, M. P. (2023). A framework for understanding label leakage in machine learning for health care. *Journal of the American Medical Informatics Association: JAMIA*, 31(1):274.

Dias, H. and Ulbrich, A. H. D. P. d. (2022). BRATECA (Brazilian Tertiary Care Dataset): a Clinical Information Dataset for the Portuguese Language. *PhysioNet*. Version 1.1.

Garriga, R., Buda, T. S., Guerreiro, J., Omaña Iglesias, J., Estella Aguerri, I., and Matić, A. (2023). Combining clinical notes with structured electronic health records enhances the prediction of mental health crises. *Cell Reports Medicine*, 4(11):101260.

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. (2000). Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220.

Johnson, A. E., Pollard, T. J., Shen, L., et al. (2016). Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035.

Knaus, W. A., Draper, E. A., Wagner, D. P., and Zimmerman, J. E. (1985). Apache ii: a severity of disease classification system. *Critical Care Medicine*, 13(10):818–829.

Le Gall, J.-R., Lemeshow, S., and Saulnier, F. (1993). A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *JAMA*, 270(24):2957–2963.

Lee, J., Yoon, W., Kim, S., et al. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Nogueira, R., Pires, R., Abonizio, H., et al. (2023). Sabiá: Portuguese large language models.

Olang, O., Mohseni, S., Shahabinezhad, A., Hamidianshirazi, Y., Goli, A., Abolghasemian, M., Shafiee, M. A., Aarabi, M., Alavinia, M., and Shaker, P. (2025). Artificial intelligence-based models for prediction of mortality in ICU patients: a scoping review. *Journal of Intensive Care Medicine*, 40(12):1240–1246.

OpenAI (2023). Gpt-4 technical report.

Seinen, T. M., Kors, J. A., van Mulligen, E. M., and Rijnbeek, P. R. (2025). Using structured codes and free-text notes to measure information complementarity in electronic health records: Feasibility and validation study. *Journal of Medical Internet Research*, 27:e66910.

Tavabi, N., Singh, M., Pruneski, J., and Kiapour, A. M. (2024). Systematic evaluation of common natural language processing techniques to codify clinical notes. *Plos one*, 19(3):e0298892.

Vagliano, I., Dormosh, N., Rios, M., Luik, T. T., Buonocore, T., Elbers, P. W., Dongelmans, D. A., Schut, M. C., and Abu-Hanna, A. (2023). Prognostic models of in-hospital mortality of intensive care patients using neural representation of unstructured text: A systematic review and critical appraisal. *Journal of Biomedical Informatics*, 146:104504.

van Aken, B., Papaioannou, J.-M., Mayrdorfer, M., Budde, K., Gers, F., and Loeser, A. (2021). Clinical outcome prediction from admission notes using self-supervised knowledge integration. In Merlo, P., Tiedemann, J., and Tsarfaty, R., editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 881–893, Online. Association for Computational Linguistics.

Vincent, J.-L., Moreno, R., Takala, J., Willatts, S., De Mendonça, A., Bruining, H., Reinhart, K., Suter, P. M., and Thijs, L. G. (1996). The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. *Intensive Care Medicine*, 22:707–710.

Do LLM Judges Agree with Humans? Evaluating Financial Commentaries from Material Facts

Gabriel Assis[1], **Daniela Vianna**[2], **Livy Real**[3,4], **Marina Ramalhete Masid**[2], **João Nepomuceno**[1], **Eduardo Rottschaefer**[1], **Altigran Soares da Silva**[3], **Aline Paes**[1]

[1] Universidade Federal Fluminense, Niterói, RJ, Brasil
{assisgabriel@id,joaonepomuceno@id,eduoliveira@id,alinepaes@ic}.uff.br,
[2]Jusbrasil, Salvador, BA, Brasil
{daniela.vianna,marina.ramalhete}@jusbrasil.com.br,
[3]Universidade Federal do Amazonas, Manaus, AM, Brasil
{livy,alti}@icomp.ufam.edu.br,
[4]Instituto Kunumi, Belo Horizonte, MG, Brasil

Abstract. *The rapid adoption of large language models (LLMs) in finance has enabled automated generation of in-domain texts such as earnings summaries, market analyses, and commentary on regulated disclosures. Generating accurate and accessible financial commentary from material facts poses challenges related to domain-specific language, strict factual faithfulness, and readability. Evaluating such outputs is difficult: traditional automatic metrics overlook financial correctness and adequacy, while human expert assessment is reliable but costly and subjective. This paper proposes a multidimensional evaluation protocol for generating financial commentary in Portuguese and investigates the use of LLMs as evaluators ("LLMs-as-judges") in this high-stakes, low-resource setting. We systematically compare human expert judgments and LLM-based evaluation preferences, while also investigating complementary dimensions such as writing quality, factuality, usefulness, and simplicity. To the best of our knowledge, this is the first proposal of an evaluation protocol for this task in Portuguese. Furthermore, we analyze where LLM judgments align with or diverge from human assessments, providing practical insights and recommendations for evaluation methodologies in financial AI applications.*

1. Introduction

Financial Commentary Generation from Material Facts is the task of producing analytical narratives from regulatory disclosures that publicly listed companies are legally required to release [Assis et al. 2025a]. "Financial commentary" refers to interpretive texts that explain, contextualize, and analyze such disclosures or related market events, ultimately helping investors make informed decisions [Zhu et al. 2023]. In practice, investors often rely on expert-written reports from financial analysts and investment research firms, motivating recent efforts to automate parts of this analytical work with language models.

Recently, the rise of large language models (LLMs) has created opportunities for financial text generation in multiple languages, including Portuguese, with applications ranging from earnings summaries [Mukherjee et al. 2022] to market analysis [Xiao et al. 2025] and investment decision support [Assis et al. 2025a]. In particular, the Brazilian financial market presents important linguistic and domain-specific challenges that remain underexplored in predominantly English-centric research. At the same

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 26–40
October 19-22, 2026 ©2026 Association for Computational Linguistics

time, finance is a high-stakes domain where errors can lead to substantial financial consequences. As a result, generated commentaries must balance readability and coherence with factual consistency, faithfulness, and domain-specific precision [Masid et al. 2026].

These requirements make the evaluation of financial commentary generation particularly challenging, as quality is multi-dimensional and often depends on specialized financial interpretation. Although traditional metrics such as BLEU [Papineni et al. 2002] and ROUGE [Lin 2004] have been previously adopted in financial text generation and related NLG tasks, their evaluation is primarily based on lexical overlap, limiting their ability to capture factual accuracy, financial validity, and deeper semantic aspects of generated commentaries. Semantic metrics such as BERTScore [Zhang et al. 2020] provide a richer semantic comparison between texts than lexical-overlap metrics, yet they may still overlook domain-specific inaccuracies or financially meaningful errors [Masid et al. 2026].

As a result, recent work has increasingly investigated the use of LLMs as evaluators [Masid et al. 2026]. The *LLM-as-a-judge* paradigm [Zheng et al. 2023, Liu et al. 2023] relies on the ability of these models to approximate human preferences while supporting scalable evaluation across multiple dimensions. In this work, we employ a *"sequential knockout pairwise evaluation"* protocol [Sandan et al. 2025], in which outputs are compared pairwise across successive evaluation rounds. This approach enables scalable analyses of evaluator agreement and robustness through comparative assessments by elimination, reducing the cost and effort associated with extensive human evaluation. However, its robustness remains unexplored in high-quality evaluation settings for Portuguese financial commentaries, where distinguishing among strong candidate outputs is especially challenging [Assis et al. 2025a, Masid et al. 2026]. In addition, LLM-based evaluators also present important limitations, such as susceptibility to fluent but factually inaccurate generations, sensitivity to prompt design, and biases toward specific linguistic styles, verbosity levels, or response structures [Chen et al. 2024, Gao et al. 2025].

In this work, we compare automatic metrics, human judgments, and LLM-based evaluations to investigate the following research questions: **RQ1.** *Under the proposed knockout tournament protocol, which LLMs emerge as the strongest generators of financial commentary?*; **RQ2.** *How do automatic metrics, LLM-based evaluations, and human assessments correlate?*; and **RQ3.** *Where do systematic disagreements arise across evaluations?*. We also analyze human annotations independently to better understand agreement patterns and the dimensions that pose the greatest challenge for expert evaluation.

To investigate these questions, our sequential knockout evaluation is inspired by prior work on pairwise preference evaluation [Sandan et al. 2025, Thakur et al. 2024] and arena-style comparisons [Zheng et al. 2023]. In this setting, commentaries are compared pairwise, with the winner advancing to the next round. Evaluators may optionally consider dimensions such as writing quality, factuality, usefulness, and simplicity when eliciting their preferences.

The strategy is instantiated using LLM4FinComm [Assis et al. 2025a], a dataset of Portuguese financial commentaries generated by both humans and LLMs. Human judgments are collected through a web-based evaluation platform[1] developed for this study, with annotations from three evaluators: a linguist and two financial-domain

[1] GitHub: https://github.com/AIDA-BR/human_and_judge_eval

experts. The same approach is further reproduced with three flagship LLMs — Qwen3-235B [Yang et al. 2025], Claude Opus 4.6 [Anthropic 2026], and Gemini 3.1 Pro [Google DeepMind 2026].

Our main contributions are threefold: (i) we propose a sequential knockout evaluation protocol for financial commentary assessment; (ii) we introduce the first benchmark for Portuguese financial commentary evaluation comparing human experts and LLM-based judges; and (iii) we provide an analysis of systematic disagreements between human and LLM evaluation preferences. The results reveal three main findings. First, GPT-4o and Llama-4-Scout emerge as strong candidates in our evaluation. Second, automatic metrics show moderate alignment with the aggregated rankings and can reliably identify strong candidates but fail to discriminate among top-performing models. Third, the LLM judges and human experts employed as evaluators in this experiment differed in meaningful ways, capturing different aspects of commentary quality.

2. Related Work

Financial text generation and evaluation have largely relied on traditional automatic metrics, particularly lexical-overlap measures such as ROUGE and its variants [Celikyilmaz et al. 2020]. Prior work spans financial report generation [Yan 2022], earnings call summarization [Mukherjee et al. 2022], and equity research report generation [Jin et al. 2026], with evaluation signals ranging from BLEU and ROUGE to semantic similarity and LLM-based components.

Recent studies also explore LLMs as judges of financial text generation [Xie et al. 2024]. Particularly, [Goldsack et al. 2025] show that such approaches correlate with expert assessments across multiple dimensions, though no single LLM consistently achieves the best correlation, and systematic biases such as positional bias and self-preference [Wataoka et al. 2024] effects remain a concern. Complementing this, evidence from the Earnings2Insights shared task [Takayanagi et al. 2025] reveals a divergence between systems that perform well under automated evaluation and those that better support human decision-making.

Research on Portuguese financial texts highlights additional challenges in transferring evaluation methodologies beyond English. [Masid et al. 2026] show that both traditional metrics and LLM-as-a-judge approaches struggle under controlled perturbations, often failing to detect subtle semantic and factual degradations. Similar findings are reported by [Assis et al. 2025a], who argue that automatic metrics lack sensitivity to fine-grained differences among outputs of comparable quality. Alternative QA-style approaches [Ribeiro et al. 2026] decompose evaluation into atomic factual checks, offering a complementary perspective for Portuguese financial summarization.

Our work complements this line of research by examining both human and LLM-based evaluation in the context of Portuguese financial commentary generation. Rather than relying on direct scoring, we investigate a successive pairwise comparison setup in which outputs are evaluated through iterative preference judgments.

3. A Multi-Round Pairwise Strategy for Evaluating Financial Commentaries

We frame evaluation as a sequence of pairwise preference judgments, rather than assigning absolute scores to individual outputs. This choice is motivated by the subjective

and context-dependent nature of financial commentary evaluation: different outputs may satisfy the same communicative goal in different ways, making absolute score calibration difficult across examples, annotators, and models. Pairwise comparison, in contrast, asks annotators to make a localized decision between two commentaries generated for the same material fact, allowing the evaluation to focus on concrete differences in writing quality, factuality, usefulness, and simplicity. Prior work further supports this design, showing that pairwise LLM-as-a-judge evaluations tend to better align with human judgments [Thakur et al. 2024], and that sequential pairwise evaluation strategies can reduce annotation costs while preserving correlation [Sandan et al. 2025].

For each material fact, N generated commentaries are organized into $N-1$ comparisons. In each comparison, the annotator, either human or LLM-based, observes the material fact and two candidate commentaries, selects the preferred one, and may optionally mark the criteria underlying the decision (writing quality, factuality, usefulness, and simplicity). The winner is carried forward as the current holder and compared with the next challenger until all commentaries have been assessed. Figure 1 provides an overview of this strategy. In the following subsections, we describe the data used (§ 3.1), the comparison criteria (§ 3.2), the human annotation process (§ 3.3), the annotation interface (§ 3.4), and the LLM-based evaluation setup (§ 3.5).

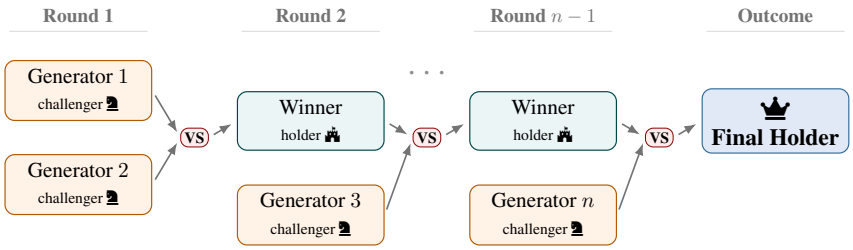


Figure 1. Multi-round pairwise evaluation under a sequential knockout strategy. Winners remain as holders and face the next challenger.

3.1. Data Selection

We use LLM4FinComm [Assis et al. 2025a] to evaluate our approach on Portuguese financial commentaries. The dataset contains 88 material fact–human-written commentary pairs. For each material fact, 9 LLM-generated commentaries are available, resulting in 792 generated texts and 880 texts in total when human references are included.

Since evaluating the full dataset would be impractical with the available human annotation resources (as discussed in § 3.3), we select a smaller subset focused on high-performing candidates. This choice aligns with RQ1, as our goal is to identify which generators are most effective for generating financial commentary. Let $\mathcal{F}$ be the set of material facts and $\mathcal{G}$ the set of LLM generators. For each pair (f, g), LLM4FinComm provides automatic scores against the human reference, including BLEU, ROUGE, BERTScore and CTC [Deng et al. 2021], a factuality metric. We select the 19 material facts with the highest-scoring associated generations. For each selected material fact, we retain the top three generated commentaries and the corresponding human-written reference, yielding four commentaries per material fact and 76 commentaries in total.

This selection strategy is also motivated by prior analyses of LLM4FinComm [Assis et al. 2025a], which show that automatic metrics have lim-

ited ability to distinguish among top-performing generators. We therefore use these metrics for pre-selection to keep the human evaluation feasible, while still enabling a later comparison between metric-based rankings and preference-based judgments. The final subset includes commentaries produced by Llama-4-Scout [Meta AI 2025]; Llama3-8B [Meta AI 2024]; Sabiá-3 [Abonizio et al. 2024]; Command-R+ [Cohere Labs 2024]; Gemma3-12B and 27B [Gemma Team 2025]; Mistral-3-24B [Mistral AI 2025]; Mistral-7B [Jiang et al. 2023]; GPT-4o [OpenAI 2024] and their human references.

3.2. Evaluation Dimensions

Annotators are presented with a set of preference dimensions that may justify their decision to support the pairwise comparison. The selection of a criterion was not obligatory, but should be done if there was an evident dimension that drove the annotator's judgment. The dimensions were defined based on prior literature on text generation evaluation and financial text assessment [Assis et al. 2025a], and then refined with the support of a linguist to improve their clarity, consistency, and adequacy for the annotation task. A pilot round of annotation was conducted to assess the usefulness of these criteria for the task. These dimensions are not used as independent scores, but as optional indicators of the aspects in which the selected commentary stands out. Thus, annotators could choose a preferred commentary whether or not they marked any criteria.

The criteria are defined as follows: (a) **Writing Quality**, the commentary is well written, clear, and remains focused on the topic introduced by the material fact; (b) **Factuality**, the commentary is faithful to the information provided in the material fact and does not introduce unsupported or misleading claims; (c) **Usefulness**, the commentary is informative and contributes concretely to the reader's understanding of the event. It may also provide relevant information about actors, entities, or contextual elements beyond those explicitly described in the material fact, when such information helps the global interpretation; and (d) **Simplicity**, the commentary uses accessible and clear language, making it understandable to non-technical readers.

In addition to these criteria, annotators are instructed to use an optional free-text field whenever they wish to explain their judgment, point out specific strengths or weaknesses, or describe relevant aspects not covered by the predefined dimensions.

3.3. Human Preference Annotation

Two annotators with complementary expertise evaluated all pairwise comparisons for each material fact, one with experience in financial contexts, and one senior linguist. Both annotators received the same comparison rounds, ensuring that disagreements could be identified directly at the level of each pairwise decision. The identity of the generators was concealed throughout the annotation process to reduce potential author-related bias and ensure that preferences were based solely on the material facts. In practical terms, this also means annotators were unaware whether each commentary came from an LLM or a human author. Additionally, in the human annotation setting, the human-written commentary was always introduced only in the final round. We adopted this strategy because, if the human commentary had been included in earlier pairwise comparisons and systematically preferred, the tournament would provide little information about which LLM-generated commentary was strongest (RQ1).

After this stage, we identified the final-round comparisons in which the annotators disagreed. These cases were adjudicated by a third senior annotator with expertise in business and finance. This adjudication step was restricted to final-round disagreements, where outcomes are most relevant for identifying the strongest candidates. Overall, this setup combines financial and linguistic perspectives while preserving blind comparisons.

3.4. Annotation Interface Design

The human annotation was conducted through a dedicated web interface developed with Streamlit[2] and designed to support the pairwise preference protocol. For each comparison round, the interface presents the material fact and two candidate commentaries side by side. After selecting a preferred commentary, annotators may optionally indicate which criteria motivated their decision using a criteria form.

3.5. LLM-as-a-Judge Annotation

The LLM-as-a-judge evaluation follows the same sequential knockout pairwise protocol used in the human annotation protocol described in § 3.3. In each comparison round, the model received the material fact and the two candidate commentaries. The prompts[3] asked the model to select the preferred commentary, indicate the criteria supporting its decision through a predefined JSON output, and provide a brief justification. As in the human setup, the identity of the generators was hidden. From the second round onward, the prompt informed the model that the current holder had won the previous comparison, mirroring the strategy's sequential structure.

Concretely, we selected three frontier models as LLM judges: Claude Opus 4.6 [Anthropic 2026], Gemini 3.1 Pro [Google DeepMind 2026], and Qwen3-235B [Yang et al. 2025]. This choice was motivated by the stronger instruction-following capabilities of recent frontier models and by their expected lower susceptibility to positional bias [Goldsack et al. 2025]. Moreover, none of these models was included among the generators evaluated, as described in § 3.1, reducing the risk of self-preference bias.

Each model was run three times using the same round configuration presented to human annotators. To assess the effect of candidate ordering, we performed three additional runs with randomized commentary order.

3.6. Evaluation and Ranking Methodology

We analyze the annotation results from two complementary perspectives. First, we adopt an Elo-based [Elo 1978] ranking strategy, following its use in Arena AI [Zheng et al. 2023], and in competitive ranking settings such as chess. Elo scores estimate a participant's relative strength by updating their rating after each comparison, giving more weight to victories over stronger opponents than to victories over weaker ones. For this analysis, we exclude the human-written commentaries, since they were always shown in the final round of the human annotation protocol, which introduces a positional bias that could distort direct Elo comparisons. After computing Elo scores for each judge, we convert them to ranks, aggregate them using the rank-sum method, and analyze the resulting correlations across judges.

[2] https://fin-eval.streamlit.app/ [3] Temperature was set to 0. The complete prompts and additional details are available at https://github.com/AIDA-BR/human_and_judge_eval

Second, we analyze agreement and disagreement patterns among LLM judges, human annotators, and the selected evaluation attributes. Rather than focusing solely on the final ranking, this perspective examines how consistently judges prefer the same outputs, how their answers correlate, and which criteria are associated with their preferences.

4. Results

This section reports and discusses the main evaluation results, covering model rankings, judge agreement, preference criteria, and comparisons with automatic metrics.

4.1. Elo-Based Ranking Analysis

Table 1 reports the rankings obtained from judge-specific Elo scores and the highest-agreement aggregated rankings, using Kendall's W as a multi-annotator agreement coefficient. Aggregated scores are computed as the rank-sum of the individual judge rankings. Considering Kendall's W for the aggregated rankings, in connection with RQ2, the strongest agreement is observed for the aggregation of the three LLM judges, followed by the combination of Gemini, the linguist annotator, and the financial expert. Among model–human pairs, Gemini is closest to the linguist annotator in the Elo ranking, suggesting that they may prioritize more similar aspects of the generated commentaries. Combinations involving LLMs and the linguist annotator also show substantial agreement in general, indicating that LLM judges tend to align more closely with the linguist annotator than with the financial expert, in line with RQ3. These results indicate that agreement patterns are not simply separated between human and LLM judges; rather, some human–LLM combinations remain consistent.

Table 1. Individual and aggregated rankings derived from Elo across judges.

Model	Individual judges					Aggregated rank					
	Q	C	G	L	F	QCG $W = .807$	GLF $W = .726$	CGL $W = .667$	QGL $W = .619$	QCL $W = .559$	All $W = .499$
GPT-4o	6 (992.75)	2 (1038.68)	**1** (1136.31)	**1** (1024.91)	3 (1002.14)	2	**1**	**1**	**1**	2	**1**
Llama 4 Scout	4 (1013.47)	6 (996.90)	2 (997.73)	2 (1009.43)	2 (1019.02)	5	2	2	2	4	2
Gemma 3 27B	**1** (1068.36)	**1** (1040.27)	4 (993.70)	7 (994.23)	8 (990.28)	**1**	7	3	3	**1**	4
Sabiá-3	3 (1023.74)	3 (1018.20)	3 (997.37)	6 (995.98)	5 (996.09)	3	5	4	5	5	3
Mistral 3 24B	2 (1051.49)	4 (1011.57)	5 (990.18)	5 (996.42)	7 (992.43)	4	6	5	4	3	5
Gemma 3 12B	5 (1000.81)	5 (998.03)	6 (988.13)	4 (997.99)	4 (998.00)	6	4	6	6	6	6
Command R+	8 (937.49)	9 (951.77)	8 (969.74)	3 (998.39)	**1** (1019.26)	8	3	7	7	7	7
Llama 3 8B	7 (977.28)	7 (988.12)	7 (976.40)	8 (994.17)	6 (994.17)	7	8	8	8	8	8
Mistral 7B	9 (934.86)	8 (956.75)	9 (951.48)	9 (988.31)	9 (988.37)	9	9	9	9	9	9

Notes. Q = Qwen, C = Claude, G = Gemini, L = linguist, and F = finance expert. Individual cells report the bootstrap rank, with the corresponding bootstrap Elo median in parentheses. Aggregated columns report ranks obtained by summing bootstrap ranks across the indicated judges; lower ranks are better.

When all judges are combined, Kendall's W is moderate (0.499), confirming disagreement in the overall ranking composition. For this reason, we do not treat any aggregated ranking as definitive. Nevertheless, with respect to RQ1, GPT-4o ranks among the top models across judges, and Llama-4-Scout ranks highly in most aggregated settings. Gemma-3 27B and Sabiá-3 also remain competitive across several settings, while Command-R+ is ranked highly by the financial expert, although this is not a general pattern across all judge types. Overall, these findings provide evidence, related to RQ1 and RQ3, that model quality depends on the evaluation perspective, with different judges emphasizing different aspects of financial commentary generation.

Table 2. Correlation between automatic metrics and the all-judge Elo ranking.

Metric combination	Spearman	Kendall τ	Top-3	Top-5
BLEU + $\text{CTC}_{\text{factual-ref}}$	**0.483**	**0.333**	0.333	**0.800**
$\text{BERTScore}_{\text{recall}}$ + $\text{BERTScore}_{\text{f1}}$ + ROUGE-1	**0.483**	**0.333**	**0.667**	**0.800**
$\text{BERTScore}_{\text{recall}}$ + ROUGE-1	0.467	0.278	**0.667**	**0.800**
BLEU + ROUGE-L + $\text{CTC}_{\text{factual-ref}}$	0.450	0.278	**0.667**	**0.800**

Addressing RQ2, Table 2 reports the ranking correlations between the automatic metrics available in LLM4FinComm and the all-judge Elo ranking. We computed all possible metric combinations and, due to space constraints, report only the four combinations with the highest correlations. Overall, the results show moderate Spearman correlations and weak Kendall correlations, suggesting that automatic metrics do not recover a full ranking that closely matches the Elo-based ordering. However, we also report the overlap with the top-3 and top-5 models in the reference ranking. Some of the best-performing metric combinations achieve reasonable coverage, reaching 0.667 overlap for the top-3 and 0.800 for the top-5. This suggests that automatic metrics may be useful for identifying strong candidates, but are less reliable for fine-grained ranking among models with similar performance [Assis et al. 2025b]. We also observe that the best-performing combinations include lexical, semantic, and factuality-oriented metrics, such as BLEU, ROUGE, BERTScore components, and CTC-based factuality metrics. This indicates that financial commentary evaluation involves complementary dimensions that no single family of automatic metrics can fully capture.

4.2. Human and LLM Judge Comparison

We use consensus across judges runs and human evaluation to assess the reliability of the sequential knockout pairwise protocol. Because the protocol is order-sensitive, we first analyze positional effects from reordering. The starting commentary wins in 28.1% of the rounds, only slightly above the random expectation of 25% for four candidates per round, suggesting that the initial position has at most a minor influence on the outcome. However, the three-run design mitigates this effect, with LLMs reaching unanimous decisions on 65–90% of items.

Humans comparison The linguist (L) and the financial expert (F) agreed on 42.1% of items (Cohen's $\kappa = 0.30$), a level of agreement that, while modest, sits well above the 17.5% expected by chance and reflects distinct evaluation stances rather than annotation noise. The third annotator, a senior financial expert, then adjudicated the remaining cases and sided with F in 72.7% of them. This divergence suggests distinct evaluation stances in our sample, with the linguist giving greater weight to communicative effectiveness and the financial experts placing more emphasis on practical utility. For example, the adjudicator sometimes cited neutrality when justifying their choice, suggesting a concern with more impartial interpretations, whereas L valued the addition of contextual information that made the text more informative. This is reflected in their choices: annotators with a financial background generally preferred LLM-generated commentaries, while L more often favored the human-written references.

The criteria profiles further support this interpretation. When F marked criteria, the selections were more often associated with usefulness (88%) and writing quality (80%) than with factuality (46%). This suggests that, for F, factual correctness may have functioned less as an explicit comparative dimension than as a prerequisite, with prefer-

ences then shaped mainly by utility and presentation quality. It is also relevant to recap that marking criteria was optional for all human annotators.

LLM Judges comparison Pairwise agreement among LLM judges, computed on the consolidated per-item winner, ranges from 36.8% to 47.4% (averaging 42.1%); on individual runs, it is somewhat higher (42.9%–49.1%, averaging 46.1%), since consolidating three runs absorbs within-judge variation. All three judges converge on the same winner in 31.6% of cases. These item-level agreement figures differ from the rank-level agreement reported in Table 1, since judges may share a similar overall ordering while still disagreeing on many individual pairwise decisions. Criterion profiles further differentiate the judges: Qwen emphasizes factuality (96%) but rarely flags writing quality (40%); Gemini often selects factuality (97%) and simplicity (73%); and Claude is more balanced, flagging factuality (96%) and utility (85%).

Human vs. LLM Judges comparison Cross-group agreement averages 27.2%, below the within-group agreement observed among both LLMs and humans (42.1% in each group). This indicates that LLM judges are not more internally consistent than human experts. At the *pairwise-decision level*, the highest human–LLM agreement is 36.8%, observed for both Claude–linguist and Gemini–financial expert. However, Gemini and the linguist show the closest Elo-based rankings (Table 1), suggesting that agreement rates and ranking similarity capture complementary aspects of alignment. In contrast, Claude shows the lowest agreement with the financial expert (21.1%).

Model preferences also differ across groups. LLM judges favor GPT-4o most often (31.6% of wins), followed by Mistral-3 and Llama-4 (15.8% each), and human texts (14.0%). Although these model outputs appeared with different frequencies due to metric-based pre-selection (§3.1), participation frequency alone does not explain preference, as Llama-4 and human texts win only 25.0% and 14.0% of their battles, respectively. Human annotators instead favor Llama-4 (28.9%) and human texts (26.3%), while also assigning votes to Command-R+ (10.5%), which receives no votes from LLM judges. Conversely, Gemma-3-27B accounts for 12.3% of LLM preferences but receives no human votes. These divergences suggest that LLM judges and human experts capture partly different aspects of commentary quality, including textual quality and decision-support value.

The variation in winners limits the use of win frequency as evidence for a single best model. Still, Llama-4-Scout emerges as a strong candidate, with consistent selections across groups, while human commentaries receive many votes, particularly from human annotators and Claude.

4.3. Automatic Metrics and Judge Preferences

The automatic metrics available in LLM4FinComm are mostly reference-based. In this sense, we can infer that the linguist (L) tends to prefer outputs with greater lexical proximity to the human reference, as reflected by the alignment with lexical automatic metrics in 81.8% of their LLM preferences. This preference may also be related to the contextual information present in human-written commentaries, which is often absent from LLM-generated outputs. The CTC-based factuality metrics also reveal an interesting dynamic. Final winners have lower CTC scores than their al- ternatives in 57.1% of comparisons overall, with this effect concentrated among LLM judges (Qwen3: 73.7%, Opus: 63.6%, Gemini: 57.9%). Human annotators do not show the same inverse pattern, with L's win-

ners having higher CTC scores in 63.6% of cases, while the financial expert (F) is roughly neutral (47.1% lower vs. 52.9% higher).

Command-R+ illustrates this dynamic. Its generations receive no votes from LLM judges, yet F ranks the model first overall, with Command-R+ accounting for 24.6% of F's preferences. F's justifications emphasize structured financial reasoning and actionable insight, dimensions that LLM judges and automatic metrics may fail to capture and that reference-based metrics may even penalize when outputs diverge from the reference. This is consistent with prior work showing that systems favored by automatic or LLM-based scores do not necessarily better support human decision-making [Takayanagi et al. 2025]. At the same time, it contrasts with some comments from the senior financial adjudicator, who often emphasized neutrality as a desirable property. Still, the adjudicator sided with F in 72.7% of disputed items, suggesting some degree of domain-aligned preference, although differences in seniority and evaluation stance may also play a role. A broader pool of annotators would be needed to better separate and understand these effects.

5. Conclusions

This paper proposes a sequential knockout protocol for evaluating financial commentaries and examines whether LLM judges approximate human expert preferences under the same evaluation framework. Our results provide evidence on three fronts:

For RQ1, on model performance, GPT-4o and Llama-4-Scout are consistently well positioned across evaluation perspectives, while Gemma-3-27B performs better under LLM judges and Command-R+ is distinctively preferred by the financial expert, despite receiving no LLM votes. For RQ2, on evaluation alignment, automatic metrics do not reproduce judge-based rankings, either human or LLM-based, contrasting with [Thakur et al. 2024]. We attribute this difference to our candidate selection strategy, which focuses on already high-scoring commentaries and makes fine-grained ranking more difficult for traditional metrics; still, automatic metrics remain useful for candidate pre-selection, covering up to 80% of top-ranked models, though not in the same order. For RQ3, on systematic disagreements, within-group agreement is comparable for LLM judges and human annotators (42.1%), whereas cross-group agreement drops to 27.2%, suggesting that disagreements are driven mainly by differences between evaluation perspectives rather than by greater instability in either group.

Finally, the sequential knockout protocol offers a practical alternative by reducing the number of comparisons required from annotators [Sandan et al. 2025]. However, deriving stable conclusions from this setup remains non-trivial in this domain. We therefore combine Elo-based ranking with agreement analysis to extract complementary signals about model performance and judge alignment. Future work should expand the annotator pool and compare this protocol with alternative designs, such as non-retained-winner or exhaustive pairwise evaluation.

Acknowledgments

This work received support from CNPq (National Council for Scientific and Technological Development) under grants 307088/2023-5, held by Aline Paes, and 301925/2025, held by Altigran Soares da Silva, and through the AIDEn Project (Proc. 400936/2025-9). From FAPERJ (Proc. SEI-260003/002930/2024 and SEI-260003/000614/2023), from

FAPEAM through the POSGRAD 2025 Program and the NeuralBond Project (grant 01.02.016301.04300/2023-04), and from CAPES – Finance Code 001. Additional support was provided by the CNPq Brazilian National Institutes of Science and Technology (INCTs), IAIA (grant #406417/2022-9), and TILD-IAR (grant #408490/2024-1). We thank Raquel da Silva Carlim for her contribution to the annotation process.

Limitations

Our analysis is based on 19 material facts, reflecting both the feasibility of annotation [Sandan et al. 2025] and our focus on the highest-scoring candidate set. This limited sample size may affect statistical precision. In addition, material facts were preselected using automated metrics, which may bias the subset toward cases with smaller metric-judge gaps than in the full dataset. In the human annotation protocol, the human-written commentary was always introduced in the final round, creating a positional confound that we partially address through Elo-based ranking and repeated LLM runs with reordered candidates. We adopt Elo as an interpretable first step for ranking pairwise preferences. Although sparse comparisons can affect Elo stability, our bootstrap analysis mitigates this limitation through bootstrap resampling. Future work may complement this analysis with order-invariant models such as Bradley–Terry–Luce or Thurstone models [Lipovetsky and Conklin 2003]. We used the same guidelines for human and LLM judges, stating that choosing a criterion for the preference was optional, but LLMs tend to point to at least one criterion, whereas humans do not. This is another point to consider when comparing LLM and human evaluation: even when using the same evaluation guidelines, they may produce different outputs that are difficult to compare. The Cohen's κ among annotators further reflects the difficulty and subjectivity of the task, as different annotators may emphasize different dimensions of commentary quality [Sharma et al. 2026]. Larger samples, more annotators, and alternative ordering strategies would strengthen generalizability, although balancing this demand with the practical constraints of expert human evaluation remains challenging, especially for senior financial expertise.

Ethics Statement

Annotation was conducted by three collaborators (one linguist and two financial-domain experts) who volunteered and were informed about the study's purpose. The dataset (LLM4FinComm) consists of publicly disclosed material facts and generated commentaries. By documenting systematic divergences between automatic, LLM-based, and expert evaluation, our work cautions against the naive use of automated metrics in financial NLP, where unreliable quality assessment may mislead downstream users, including investors.

Disclosure of GenAI usage

The LLMs evaluated in this study are described in §3. GenAI tools (Claude Code and GPT Codex) were used to support auxiliary data-processing tasks via Python scripts, including data preparation, file manipulation, and CSV formatting. GenAI tools were also used for language polishing, translation between Portuguese and English, and LaTeX coding. All generated or tool-assisted content was reviewed by the authors, who remain responsible for the analyses, interpretations, and conclusions presented in this work.

References

Abonizio, H., Almeida, T. S., Laitz, T., Junior, R. M., Bonás, G. K., Nogueira, R., and Pires, R. (2024). Sabiá-3 technical report.

Anthropic (2026). Claude opus 4.6 system card. https://www-cdn.anthropic.com/14e4fb01875d2a69f646fa5e574dea2b1c0ff7b5.pdf.

Assis, G., Dutra, H., Vianna, D., Meira, Wagner, J., da Silva, A. S., and Paes, A. (2025a). Language models for automated market commentary from corporate disclosures. In *Proceedings of the 6th ACM International Conference on AI in Finance*, ICAIF '25, page 727–735, New York, NY, USA. Association for Computing Machinery.

Assis, G., Freitas, C., and Paes, A. (2025b). Exploring Brazil's LLM Fauna: Investigating the generative performance of large language models in portuguese. *Journal of the Brazilian Computer Society*, 31(1):939–971.

Celikyilmaz, A., Clark, E., and Gao, J. (2020). Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Chen, G. H., Chen, S., Liu, Z., Jiang, F., and Wang, B. (2024). Humans or LLMs as the judge? a study on judgement bias. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, Miami, Florida, USA. Association for Computational Linguistics.

Cohere Labs (2024). Introducing command r+: A scalable llm built for business. https://cohere.com/blog/command-r-plus-microsoft-azure.

Deng, M., Tan, B., Liu, Z., Xing, E., and Hu, Z. (2021). Compression, transduction, and creation: A unified framework for evaluating natural language generation. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7580–7605, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Elo, A. E. (1978). *The Rating of Chessplayers, Past and Present*. Arco Publishing.

Gao, M., Hu, X., Yin, X., Ruan, J., Pu, X., and Wan, X. (2025). LLM-based NLG Evaluation: Current Status and Challenges. *Computational Linguistics*, 51(2):661–687.

Gemma Team (2025). Gemma 3 Technical Report.

Goldsack, T., Wang, Y., Lin, C., and Chen, C.-C. (2025). From facts to insights: A study on the generation and evaluation of analytical reports for deciphering earnings calls. In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D., and Schockaert, S., editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10576–10593, Abu Dhabi, UAE. Association for Computational Linguistics.

Google DeepMind (2026). Gemini 3.1 pro model card. https://deepmind.google/models/model-cards/gemini-3-1-pro/. Accessed: 2026.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-

A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2023). Mistral 7b.

Jin, S., Li, S., Zhang, S., and Yan, R. (2026). Finrpt: Dataset, evaluation system and LLM-based multi-agent framework for equity research report generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(1):507–515.

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. ACL.

Lipovetsky, S. and Conklin, M. (2003). Priority estimations by pair comparisons: Ahp, thurstone scaling, bradley-terry-luce, and markov stochastic modeling. In *Proceedings of the ASA Joint Statistical Meeting, JSM*.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. (2023). G-eval: NLG evaluation using gpt-4 with better human alignment. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Masid, M. R., Assis, G., Vianna, D., Paes, A., and Silva, A. S. d. (2026). Auditing the evaluators: How far can automatic evaluation go in assessing Portuguese financial texts? In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 222–233, Salvador, Brazil. Association for Computational Linguistics.

Meta AI (2024). The llama 3 herd of models.

Meta AI (2025). The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation.

Mistral AI (2025). Mistral Small 3.1.

Mukherjee, R., Bohra, A., Banerjee, A., Sharma, S., Hegde, M., Shaikh, A., Shrivastava, S., Dasgupta, K., Ganguly, N., Ghosh, S., and Goyal, P. (2022). ECTSum: A new benchmark dataset for bullet point summarization of long earnings call transcripts. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10893–10906, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

OpenAI (2024). Gpt-4o system card.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on ACL*, ACL '02, page 311–318, USA. ACL.

Ribeiro, J. V. A., Correia, T. P., Requena, J. V. S. C., and Berton, L. (2026). Evaluating reference-free summarization quality metrics for Portuguese: A study with human judgments in financial news. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 899–907, Salvador, Brazil. Association for Computational Linguistics.

Sandan, I. B., Dinh, T. A., and Niehues, J. (2025). Knockout LLM assessment: Using large language models for evaluations through iterative pairwise comparisons. In Ar-

viv, O., Clinciu, M., Dhole, K., Dror, R., Gehrmann, S., Habba, E., Itzhak, I., Mille, S., Perlitz, Y., Santus, E., Sedoc, J., Shmueli Scheuer, M., Stanovsky, G., and Tafjord, O., editors, *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 121–128, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Sharma, N., Agarwal, N., and Sirts, K. (2026). Towards consistent detection of cognitive distortions: Llm-based annotation and dataset-agnostic evaluation. In Piperidis, S., Bel, N., van den Heuvel, H., Ide, N., Krek, S., and Toral, A., editors, *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 10866–10882, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Takayanagi, T., Goldsack, T., Izumi, K., Lin, C., Takamura, H., and Chen, C.-C. (2025). Earnings2Insights: Analyst report generation for investment guidance. In Chen, C.-C., Winata, G. I., Rawls, S., Das, A., Chen, H.-H., and Takamura, H., editors, *Proceedings of The 10th Workshop on Financial Technology and Natural Language Processing*, pages 246–251, Suzhou, China. Association for Computational Linguistics.

Thakur, A. S., Choudhary, K., Ramayapally, V. S., Vaidyanathan, S., and Hupkes, D. (2024). Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics*.

Wataoka, K., Takahashi, T., and Ri, R. (2024). Self-preference bias in LLM-as-a-judge. In *Neurips Safe Generative AI Workshop 2024*.

Xiao, Y., Sun, E., Luo, D., and Wang, W. (2025). Tradingagents: Multi-agents LLM financial trading framework. In *The First MARW: Multi-Agent AI in the Real World Workshop at AAAI 2025*.

Xie, Q., Han, W., Chen, Z., Xiang, R., Zhang, X., He, Y., Xiao, M., Li, D., Dai, Y., Feng, D., et al. (2024). Finben: A holistic financial benchmark for large language models. *Advances in Neural Information Processing Systems*, 37:95716–95743.

Yan, S. (2022). Disentangled variational topic inference for topic-accurate financial report generation. In Chen, C.-C., Huang, H.-H., Takamura, H., and Chen, H.-H., editors, *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 18–24, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., and Qiu, Z. (2025). Qwen3 technical report.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., Zhang, H., Gonzalez, J., and Stoica, I. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Zhu, D., Lappas, T., and Rachidi, T. (2023). Commentary generation for financial markets. *Expert Systems with Applications*, 211:118364.

Expressões multipalavras no discurso de fóruns sobre saúde mental: desenvolvimento de um *corpus* em inglês e português brasileiro

Gabriela Berndt[1], Sofia Serrano[1], Mariana Jin [1], Adriana Pagano[1]

[1] Universidade Federal de Minas Gerais
{gabiberndt1, sofias, mjin, apagano}@ufmg.br

Abstract. *Multiword expressions (MWEs), word sequences displaying some degree of idiosyncrasy, are central to both linguistic studies and Natural Language Processing (NLP). This paper reports on the creation of a novel corpus of MWEs in English and Brazilian Portuguese in posts from Reddit two Reddit threads about depression — a domain underrepresented in existing MWE resources, predominantly based on newspaper texts. The corpus was manually compiled, automatically parsed using UDPipe, revised in Arborator Grew and manually annotated on the FLAT platform following the PARSEME guidelines. The analysis combines frequency distribution by MWE category and cross-linguistic contrastive examination. The results highlight domain-specific patterns in MWEs, as well as constructions that function as conventionalized resources for expressing feelings, evaluations, and subjective states in digital mental health discourse.*

Resumo. *Expressões multipalavras (multiword expressions, MWEs), isto é, sequências de palavras que apresentam algum grau de idiossincrasia, são centrais tanto para os estudos linguísticos quanto para o Processamento de Linguagem Natural (PLN). Este artigo apresenta a criação de um novo corpus de MWEs em postagens em inglês e português brasileiro de comunidades do Reddit dedicadas à saúde mental — um domínio ainda sub-representado nos recursos existentes de MWEs, predominantemente baseados em textos jornalísticos. O corpus foi compilado manualmente, anotado automaticamente com informações morfossintáticas e sintáticas por meio do parser UDPipe, revisado no Arborator Grew e anotado manualmente na plataforma FLAT, seguindo as diretrizes do PARSEME. A análise combina a distribuição de frequência por categoria de MWE e uma investigação contrastiva interlinguística. Os resultados evidenciam especificidades das MWEs nesse domínio e construções que funcionam como recursos convencionalizados para expressar sentimentos, avaliações e estados subjetivos no discurso sobre saúde mental nas redes sociais.*

1. Introdução

As expressões multipalavras (MWEs) constituem um fenômeno linguístico de grande relevância para o processamento de linguagem natural (PLN), uma vez que representam unidades lexicais complexas cujo significado é não-composicional, isto é, não derivável do significado individual de seus constituintes [Firth 1957], [Ramisch 2015].Essas combinações apresentam idiossincrasias sintáticas e semânticas que dificultam sua

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 41–49
October 19-22, 2026 ©2026 Association for Computational Linguistics

interpretação automática, impactando tarefas como análise de sentimentos, tradução automática e outras aplicações em PLN. No domínio da saúde mental, a identificação e interpretação dessas expressões é particularmente importante, pois MWEs são frequentemente utilizadas para expressar diferentes graus de intensidade afetiva, estratégias de *coping* e estados dissociativos. [Coll-Florit 2021], ao estudar metáforas, observou que em ambientes como fóruns online, essas construções – as quais são compostas por e/ou atuam como modificadores – funcionam como recursos convencionados de significação, frequentemente mediando a construção discursiva da experiência psicológica.

Apesar da centralidade das MWEs, persistem lacunas importantes na literatura: os recursos existentes para tratamento das MWEs em português brasileiro concentram-se predominantemente em textos jornalísticos, de modo que não contemplam textos produzidos em redes sociais, classificados como vinculados à atividade sócio-semiótica compartilhar (*sharing*), segundo a classificação de [Matthiessen 2010]. A mesma lacuna se observa para o inglês no domínio específico de fóruns digitais. Em ambas as línguas, ainda, textos relacionados ao domínio da saúde mental são pouco contemplados. Embora não trate de MWEs nem inclua dados em inglês, [Wilkens et al. 2026]analisam postagens sobre saúde mental em português brasileiro, mostrando que padrões lexicais, sintáticos e psicolinguísticos distintivos podem ser associados a perfis de usuários com depressão. Essa perspectiva está em sintonia com o presente estudo, que investiga as MWEs como recursos convencionalizados para expressar sentimentos, avaliações e estados subjetivos no discurso digital sobre saúde mental.

Diante do cenário descrito, este trabalho teve como objetivo principal a criação de um *dataset* composto por dois subcorpora, um em português brasileiro e outro em inglês, de textos extraídos de fóruns digitais do Reddit sobre saúde mental, anotado com categorias de MWEs de acordo com o manual de anotação do projeto PARSEME. Com base no dataset, analisamos a distribuição tipológica das MWEs e as diferenças de uso dessas expressões entre o inglês e o português brasileiro. Embora ainda seja de dimensão reduzida, o *dataset* permitiu aplicar e refinar um protocolo de anotação para os objetivos da pesquisa. O corpus poderá apoiar a avaliação e a análise de erros de sistemas de identificação de MWEs.

2. Referencial Teórico e Fundamentos da Anotação

2.1. Definição de MWEs

MWEs são construções gramaticais complexas passíveis de ser analisadas em mais de um lexema e caracterizadas pela não composicionalidade [Baldwin and Kim 2010], [Constant et al. 2021], [Ramisch and Villavicencio 2018]. O elemento comum entre as definições consolidadas na área [Calzolari et al. 2002], [**?**] é a ideia de que o significado do conjunto não é derivado da análise das partes e nem das regras morfossintáticas que as combinam. Trata-se, então, de uma questão de grau: as MWEs podem ser mais ou menos composicionais [Cordeiro et al. 2019], distribuídas em um *continuum* que vai de construções transparentes a expressões completamente opacas.

2.2. O Projeto PARSEME e a Taxonomia das MWEs

No campo do PLN, a necessidade de sistematização levou ao desenvolvimento de tipologias operacionais de MWEs. Nesse contexto, destaca-se o projeto PARSEME

[Savary et al. 2017, Ramisch et al. 2020], [Savary et al. 2023], [Scholivet et al. 2026], [Savary et al. 2026], que propõe uma taxonomia padronizada com categorias voltadas à anotação multilíngue de MWEs, detalhada por [Savary et al. 2023] e [Savary et al. 2017]. Essa tipologia tornou-se referência para análises comparativas e para a construção de *corpus* anotados em múltiplas línguas, sendo adotada integralmente neste trabalho. As denominações em português dessas categorias são traduções/adaptações nossas, com base nas diretrizes PARSEME 2.0 ([Savary et al. 2026] e na taxonomia discutida por [Savary et al. 2023], retomada no capítulo de [Ramisch et al. 2024]. Neste estudo, foram consideradas 15 categorias da taxonomia adotada: expressões idiomáticas verbais (*verbal idioms*, VID); construções verbo-partícula (*idiomatic verb-particle constructions*, IVPC); construções verbo-suporte (*light-verb constructions*, LVC); verbos inerentemente reflexivos (*inherently reflexive verbs*, IRV); construções multiverbais (*multi-verb constructions*, MVC); expressões verbais inerentemente adposicionais (*inherently adpositional verbs*, IAV); expressões idiomáticas nominais (*nominal idioms*, NID); expressões idiomáticas pronominais (*pronominal idioms*, PronID); expressões nominais deverbais (*deverbal nominal idioms*, NV) (Ramisch et al. 2026); expressões idiomáticas adjetivais (*adjectival idioms*, AdjID); expressões idiomáticas adverbiais (*adverbial idioms*, AdvID); expressões idiomáticas determinantes (*determiner idioms*, DetID); expressões idiomáticas adposicionais (*adposition idioms*, AdpID); expressões idiomáticas conjuntivas (*conjunction idioms*, ConjID); e expressões idiomáticas interjeitivas (*interjection idioms*, IntjID). As categorias efetivamente identificadas e suas frequências são apresentadas na Tabela 3..

Este estudo examina as MWEs em um *corpus* comparável bilíngue, articulando duas dimensões analíticas: a distribuição tipológica das expressões, com base na classificação do PARSEME, e seus padrões de uso em discurso digital sobre saúde mental. A partir dessa abordagem, examinam-se a distribuição tipológica das MWEs e seus padrões de funcionamento discursivo nos dois subcorpus, com atenção às convergências e divergências entre o inglês e o português brasileiro no discurso digital sobre saúde mental.

3. *Corpus* e Metodologia

3.1. Composição do *Corpus*

O *corpus* constitui-se de textos coletados na plataforma Reddit, escolhida por sua natureza multilíngue e pelo caráter espontâneo e informal das interações, que propicia um discurso indireto e contextualizado [Wilkens et al. 2026] e que favorece a emergência de MWEs em contextos autênticos de uso. O *subcorpus* em português foi extraído da comunidade r/perguntas, a partir do tópico *O que vocês fazem para lidar com a depressão*; o *subcorpus* em inglês foi formado a partir de respostas ao tópico *how would you describe depression to someone who's never experienced it?/como você descreveria a depressão para alguém que nunca passou por isso?*, da comunidade r/AskReddit. As perguntas não são equivalentes: a pergunta em português solicita práticas utilizadas para lidar com a depressão, enquanto a pergunta em inglês solicita uma descrição da experiência. Essa diferença pode influenciar a distribuição das categorias e impede que os contrastes sejam atribuídos exclusivamente às línguas. Nos dois casos, privilegiaram-se ambientes discursivos em que as pessoas descrevem, explicam ou elaboram suas vivências de sofrimento psíquico. Os dados foram coletados exclusivamente de fontes públicas, com total supressão de metadados de identificação, em conformidade com boas práticas de pesquisa ética em *corpus*. O pré-processamento incluiu normalização ortográfica, expansão de siglas, remoção de

metadados e *emojis* não interpretáveis e segmentação manual de comentários extensos em unidades sentenciais. Sempre que possível, foi preservada a segmentação espontânea realizada pelos próprios autores dos comentários. De forma que apenas em casos de comentários extensos e sem pontuação ortográfica clara realizou-se segmentação manual, tomando-se as orações como unidades de significado.

3.2. Caracterização do corpus

A Tabela 1 apresenta uma caracterização geral do *corpus* nas duas línguas analisadas-quanto ao número de sentenças e palavras e à razão entre esses parâmetros. Os dados apresentados são aqui expostos para explicitar o escopo do *corpus* assim como demonstrar que os *subcorpora* são equiparáveis, o que também é verdade para a Tabela 2.

Tabela 1. Caracterização do *corpus*

Parâmetro	PB	EN	Razão PB/EN
Sentenças	263	261	1,01
Palavras	4.253	4.166	1,02

3.3. *Pipeline* de Anotação

O *pipeline* de anotação abrangeu duas etapas sequenciais. Na primeira, os *corpora* foram anotados morfossintaticamente por meio do parser UDPipe [Straka et al. 2016], utilizando o *corpus* Portuguese-Porttinari, identificado na plataforma como portuguese-porttinari-ud-2.17-251125, treinado a partir do treebank Porttinari-Base-ud [Duran et al. 2023] e o modelo English-GUM, identificado como english-gum-ud-2.17-251125 treinado a partir do treebank english-gum-ud [Zeldes 2017], gerando arquivos CoNLL-U. Na segunda etapa, os arquivos foram importados para a plataforma FLAT [Van Gompel et al. 2024] e as MWEs identificadas e classificadas manualmente por três anotadoras especializadas, sentença por sentença, nas duas línguas. Os critérios de identificação de MWEs contemplaram os testes propostos pelo PARSEME. A adjudicação foi conduzida por uma especialista sênior, que resolveu os casos de discordância entre as anotadoras.

4. Resultados

A Tabela 2 apresenta o número de ocorrências de MWEs e de tipos lematizados distintos; por exemplo, *fazer terapia* e *faço terapia* constituem duas ocorrências de um mesmo tipo lematizado.

Tabela 2. MWEs encontradas

Parâmetro	PB	EN	Razão PB/EN
MWEs anotadas	122	81	1,53
MWEs por lema	97	71	1,37

4.1. Distribuição por Categoria PARSEME

A Tabela 3 apresenta as categorias do PARSEME com ocorrências em cada *subcorpus* analisado (português e inglês). São apresentados o número de ocorrências e frequência relativa de cada categoria no total de ocorrências. A Tabela apresenta somente as categorias efetivamente encontradas no corpus.

Tabela 3. Categorias das MWEs nos *subcorpus*

Categoria	PB (n)	PB (%)	EN (n)	EN (%)
AdjID	3	2,46%	0	0,00%
AdpID	5	4,10%	0	0,00%
AdvID	25	20,49%	12	14,81%
ConjID	7	5,74%	4	4,94%
DetID	2	1,64%	3	3,70%
IAV	9	7,38%	7	8,64%
IntjID	3	2,46%	1	1,23%
IRV	11	9,02%	0	0,00%
IVPC	0	0,00%	31	36,71%
LVC	22	18,03%	5	6,17%
MVC	1	0,82%	1	1,23%
NID	19	15,57%	5	6,33%
NV-LVC	0	0,00%	1	1,23%
PronID	0	0,00%	1	1,23%
VID	15	12,30%	10	12,35%
Total	**122**	**100,00%**	**81**	**100,00%**

Considerando as 15 categorias analisadas anotadas na Tabela 3, 12 categorias apresentaram ao menos uma ocorrência no *subcorpus* em português brasileiro e 12 no *subcorpus* em inglês, assim como explicitado na Tabela 3. A análise feita apontou as categorias verbais como as mais numerosas nas duas línguas (47,5% e 66,67% no português e inglês) seguidas da etiquetas adverbial/adjetiva (22,95% e 14,81%) o que está de acordo com o esperado na literatura [Wilkens et al. 2026]. Os perfis de distribuição de categorias, no entanto, são distintos entre as duas línguas e essa diferença constitui o achado central deste estudo.

Três grandes diferenças se destacam e são todas teoricamente motivadas. A primeira, e mais expressiva, diz respeito às construções verbo partícula (idiomatic verb-particle constructions, IVPC): no inglês, essa categoria representa 38,27% das MWEs, enquanto ela não aparece no português. Essa divergência reflete uma propriedade tipológica bem documentada: o inglês é uma língua de enquadramento por satélite (*satellite-framed*), que expressa trajetória e aspecto por meio de partículas verbais separáveis, como ”give up”, ”break down”e ”fall apart”. Já o português, como língua de enquadramento verbal, codifica esses conteúdos no próprio verbo ou em construções perifrásticas [Talmy 2000]. No *subcorpus* analisado, essa tendência aparece em expressões verbais inerentemente adposicionais como ”lidar com”, ”contar com”, bem como em construções verbo-suporte como ”fazer terapia”, ”tomar decisão”.

A segunda diferença diz respeito aos verbos inerentemente reflexivos (inherently

reflexive verbs, IRV): o português registra 11 ocorrências (9,02%), enquanto o inglês não apresenta nenhuma instância dessa categoria. Esse resultado é compatível com diferenças estruturais entre as línguas e com os critérios da categoria PARSEME, uma vez que o inglês não dispõe de um sistema de clíticos reflexivos gramaticalmente obrigatório comparável ao do português, em que construções como ”queixar-se”e ”sentir-se”o verbo é morfossintática- e semanticamente indissociável do pronome reflexivo ”se”.

A terceira diferença está relacionada às construções verbo-suporte (light-verb constructions, LVC). Ambas as línguas se mostram produtivas, mas o português apresenta frequência proporcionalmente superior (18,03% vs. 6,17%). Em ambas as línguas, foram mapeados os verbos suporte mais frequentes (”ter”, ”tomar”, ”fazer”, ”estar”, ”dar”, em português; e ”take”, ”make”, ”get”em inglês). A razão da maior incidência de LVC no *subcorpus* do português precisa ser mais bem investigada.

Além da distribuição tipológica, os dados revelam que determinadas MWEs funcionam como recursos convencionados para a expressão de experiências emocionais e psicológicas no discurso digital sobre depressão [Coll-Florit 2021]. As MWEs do *corpus* organizam-se em torno de três funções discursivas complementares. A mais frequente é a nomeação metafórica de estados depressivos por NID e VID — expressões como ”fundo do poço”e ”rest in peace”, que a comunidade emprega como convenções estáveis de significação, com opacidade semântica parcialmente neutralizada pelo contexto discursivo compartilhado. A essa se somam os AdvID na função de calibradores da experiência subjetiva (”às vezes”, ”out of nowhere”), sinalizando frequência, imprevisibilidade e cronicidade dos estados emocionais. Por fim, as LVC constituem um vocabulário fraseológico de *coping*: construções como ”fazer terapia”e ”seek help”nomeiam práticas terapêuticas e de enfrentamento recorrentes no discurso de compartilhamento em saúde mental.

5. Conclusão

Este trabalho apresentou um corpus-piloto anotado segundo o PARSEME 2.0 e descreveu a distribuição das MWEs em dois tópicos do Reddit. O recurso poderá apoiar a avaliação e a análise de erros de sistemas de identificação de MWEs. No domínio da saúde mental digital, de crescente relevância para o PLN clínico, a diferença é especialmente significativa, reforçando a necessidade de *corpus* anotados manualmente e de recursos específicos para o português [Ramisch et al. 2024], [Savary et al. 2023]. O *corpus* compilado é de tamanho moderado e restrito a dois tópicos do Reddit, sendo assim, a generalização dos dados obtidos para outros registros e plataformas exige cautela.

6. Agradecimentos

As autoras agradecem os pareceristas anônimos pelas sugestões para o aprimoramento deste trabalho. Gabriela Berndt é bolsista de mestrado da CAPES (88887.204699/2025-00). Sofia Serrano é bolsista de monitoria de graduação da Pró-Reitoria de Graduação da UFMG. Adriana Pagano é Bolsista de Produtividade em Pesquisa e recebe apoio do CNPq (404722/2024-5; 313103/2021-6) e da FAPEMIG. Este trabalho recebeu apoio da Ação COST CA21167 UniDive, financiada pela European Cooperation in Science and Technology. A pesquisa afilia-se ao Instituto Nacional de Ciência e Tecnologia (INCT) em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação (TILDIAR), financiado pelo CNPq (408490/2024-1).

7. Limitações

Como já apontado anteriormente, o tamanho dos *corpus* deste trabalho não é suficiente para a sua utilização em *fine tuning* de grandes modelos de linguagem (LLMs). Quanto à análise dos dados, uma outra limitação é que não foi feita uma comparação com um *corpus* anotado de referência em outro domínio, em decorrência de restrições de tempo. Apesar disso, esse trabalho está incluso em projeto maior de pesquisa, ainda em andamento, o qual se debruçará em algumas lacunas além de expandir o *dataset* com um *subcorpus* em francês canadense.

8. Declaração Ética

Este estudo utiliza exclusivamente dados publicamente acessíveis na plataforma Reddit, em conformidade com seus Termos de Uso. Reconhecendo a sensibilidade do domínio de saúde mental, adotaram-se as seguintes salvaguardas: (i) anonimização integral, com remoção de nomes de usuário e identificadores de fóruns; (ii) disponibilização do *corpus* apenas em formato CoNLL-U, sem trechos suficientes para reconstituição dos posts originais; (iii) ausência de qualquer inferência sobre estados clínicos de usuários individuais. Desencoraja-se o uso do *corpus* para fins de triagem ou diagnóstico clínico.

9. Uso de Ferramentas de IA Generativa

As autoras declaram o uso de ferramentas de IA generativa para a organização e limpeza das tabelas sem a alteração desses dados ou *tags* atribuídas. Todo o conteúdo gerado foi revisado, validado e adaptado pelas autoras, que assumem responsabilidade pelo resultado final. Não foram utilizadas ferramentas de IA generativa na anotação do *corpus*, na análise dos dados ou na redação do trabalho.

Referências

Baldwin, T. and Kim, S. N. (2010). Multiword expressions. In Indurkha, N.andDamerau, F. J., editor, *Handbook of Natural Language Processing*. CRC Press.

Calzolari, N., Fillmore, C. J., Grishman, R., Ide, N., Lenci, A., MacLeod, C., and Zampolli, A. (2002). Towards best practice for multiword expressions in computational lexicons. In González Rodríguez, M. and Suarez Araujo, C. P., editors, *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. uropean Language Resources Association (ELRA).

Coll-Florit, M.; Oliver, A. R. S. (2021). Metaphors of mental illness: a corpus-based approach analysing first-person accounts of patients and mental health professionals. *Cultura, Lenguaje y Representación*, 25:85–104.

Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., and Todirascu, A. (2021). Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Cordeiro, S., Villavicencio, A., Idiart, M., and Ramisch, C. (2019). Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*, 45(1):1–57.

Duran, M., Lopes, L., das Graças Nunes, M., and Pardo, T. (2023). The dawn of the porttinari multigenre treebank: Introducing its journalistic portion. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 115–124, Porto Alegre, RS, Brasil. SBC.

Firth, J. R. (1957). *A Synopsis of Linguistic Theory*, pages 1–32. Blackwell, Oxford.

Matthiessen, C. M. I. M.; Teruya, K. L. M. (2010). *Key Terms in Systemic Functional Linguistics*. Continuum, London.

Ramisch, C. (2015). *Multiword Expressions Acquisition: A Generic and Open Framework*. springer.

Ramisch, C., Savary, A., Guillaume, B., Waszczuk, J., Candito, M., Vaidya, A., Barbu Mititelu, V., Bhatia, A., Iñurrieta, U., Giouli, V., Güngör, T., Jiang, M., Lichte, T., Liebeskind, C., Monti, J., Ramisch, R., Stymne, S., Walsh, A., and Xu, H. (2020). Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*.

Ramisch, C. and Villavicencio, A. (2018). Computational treatment of multiword expressions. In Mitkov, R., editor, *The Oxford Handbook of Computational Linguistics*. Oxford University Press, 2nd edition. `https://doi.org/10.1093/oxfordhb/9780199573691.013.56`.

Ramisch, R., Ramisch, C., and Villavicencio, A. (2024). Expressões multipalavras: fogo de palha ou osso duro de roer? In Caseli, H. and das Graças Volpe Nunes, M., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, chapter 5. BPLN, 3 edition. `https://brasileiraspln.com/livro-pln/3a-edicao/parte-palavras/cap-mwe/cap-mwe.html`.

Savary, A., Ben Khelil, C., Ramisch, C., Giouli, V., Barbu Mititelu, V., Hadj Mohamed, N., Krstev, C., Liebeskind, C., Xu, H., Stymne, S., Güngör, T., Pickard, T., Guillaume, B., Bejček, E., Bhatia, A., Candito, M., Gantar, P., Iñurrieta, U., Gatt, A., Kovalevskaite, J., Lichte, T., Ljubešić, N., Monti, J., Parra Escartín, C., Shamsfard, M., Stoyanova, I., Vincze, V., and Walsh, A. (2023). PARSEME corpus release 1.3. In Bhatia, A., Evang, K., Garcia, M., Giouli, V., Han, L., and Taslimipoor, S., editors, *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Dubrovnik, Croatia. Association for Computational Linguistics.

Savary, A., Ramisch, C., Cordeiro, S., Sangati, F., Vincze, V., QasemiZadeh, B., Candito, M., Cap, F., Giouli, V., Stoyanova, I., and Doucet, A. (2017). The PARSEME shared task on automatic identification of verbal multiword expressions. In Markantonatou, S., Ramisch, C., Savary, A., and Vincze, V., editors, *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47, Valencia, Spain. Association for Computational Linguistics.

Savary, A., Scholivet, M., Ramisch, C., Nakamura, T., Bilinski, E., Stymne, S., Giouli, V., Markantonatou, S., Pais, V., Mitrofan, M., Estève, L., Guillaume, B., Mititelu, V. B., Čibej, J., Hernández, R. D., Fendel, V., Gantar, P., Kanishcheva, O., Krstev, C., Liebeskind, C., Lobzhanidze, I., Marković, A. M., Nešpore-Bērzkalne, G., Pagano, A. S., Shamsfard, M., Stankovic, R., Tajalli, V., Tiberius, C., and Padhye, A. (2026). Parseme

2.0 multilingual corpus of multiword expressions. In Piperidis, S., Bel, N., van den Heuvel, H., Ide, N., Krek, S., and Toral, A., editors, *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 4819–4834, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Scholivet, M., Savary, A., Ramisch, C., Bilinski, E., Nakamura, T., Mitrofan, M., and Pais, V. (2026). Edition 2.0 of the PARSEME shared task on multilingual identification and paraphrasing of multiword expressions. In Ojha, A. K., Mititelu, V. B., Constant, M., Stoyanova, I., Doğruöz, A. S., and Rademaker, A., editors, *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, pages 254–275, Rabat, Marocco. Association for Computational Linguistics.

Straka, M., Hajič, J., and Straková, J. (2016). UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA).

Talmy, L. (2000). *Toward a Cognitive Semantics*. MIT Press,.

Van Gompel, M., Başar, E., Neumann, A., and van der Klis, M. (2024). Proycon/flat: V0.11.5.

Wilkens, R., Caseli, H., Neris, V., and Villavicencio, A. (2026). The visible and the latent linguistic clues of mental health in Brazilian Portuguese textual posts. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 2*, pages 135–147, Salvador, Brazil. Association for Computational Linguistics.

Zeldes, A. (2017). The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Text segmentation with mixed data: evaluating symbolic and neural approaches

Amanda Sammer do N. Brandão[1], Felipe Bomfim Sanches[1], Alfredo Sena[1], Laila Mota [1], Daniela Barreiro Claro[1],Rerisson Cavalcante[2]

[1]FORMAS - Research Center on Data and Natural Language
Institute of Computing – Federal University of Bahia (UFBA)
{amandabrandao,felipe.sanches,alfredosena,laila.pereira,dclaro}@ufba.br
[2]FORMAS - Research Center on Data and Natural Language
Institute of Letters – Federal University of Bahia (UFBA)
Salvador – Bahia – Brazil
rerisson.cavalcante@ufba.br

Abstract. *Text segmentation has been widely adopted and used to extract relevant information from documents. This information is distributed across various types of structures, including heterogeneous data. The extraction of such data has recently been extensively leveraged by Large Language Models (LLMs). However, the training of these models does not inherently capture the particular characteristics of mixed data, such as lexical and phonetic information. As a result, automatically assessing the correctness of transcriptions using LLMs becomes a challenging task. In this study, text segmentation and information extraction were parallelized into two main approaches in order to compare a symbolic approach with a neural one. The results highlight the advantages and disadvantages of both approaches when dealing with mixed data types.*

Resumo. *A segmentação de texto vem sendo amplamente difundida e utilizada para extrair informação relevante de documentos. Essas informações estão distribuídas em diversos tipos de estruturas, que incluem dados heterogêneos. A extração desses dados tem sido ultimamente, utilizada por LLMs (Large Language Models). Porém, os treinamentos dos modelos de linguagem não detém características peculiares de dados mistos, tais como dados lexicais e dados fonéticos. Assim, torna-se um desafio analisar a corretude das transcrições de maneira automatizada por LLMs. Neste estudo, a segmentação textual e extração de informação foi paralelizada em duas principais abordagens com o intuito de avaliar a comparação de uma abordagem simbólica e uma abordagem neural. Os resultados demonstram as vantagens e desvantagens em ambas as abordagens para os tipos de dados mistos.*

1. Introduction

Text segmentation and information extraction are fundamental Natural Language Processing (NLP) tasks for transforming unstructured or semi-structured documents into computationally manipulable representations [Ghinassi et al. 2024, Lu et al. 2022]. In this study, segmentation refers specifically to identifying question blocks, speech turns, and their corresponding textual content in a Phonetic-Phonological Questionnaire (QFF). Such documents combine heterogeneous linguistic data, including graphematic text, phonetic transcription, numerical question identifiers, informant responses,

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 50–61
October 19-22, 2026 ©2026 Association for Computational Linguistics

interviewer interventions, and speech-turn markers. Studies involving Brazilian Portuguese corpora highlight the importance of careful organization and manual validation [Candido Junior et al. 2023], while [Queiroz et al. 2023] demonstrate the relevance of explicit extraction rules, manual annotation, and expert validation in the construction of reliable linguistic resources. Phonetic transcriptions also require the preservation of symbols from the International Phonetic Alphabet (IPA) [International Phonetic Association 1999].

Recent advances have expanded the use of Large Language Models (LLMs) in segmentation and information extraction tasks [Krassovitskiy et al. 2025, Cabral et al. 2024]. However, the task investigated here requires more than semantic equivalence: question identifiers, speaker attribution, speech-turn boundaries, textual content, and phonetic symbols must be preserved with minimal structural modification. Because LLM outputs are probabilistic, plausible responses may not exactly reproduce the original structure, particularly regarding instruction following, schema adherence, and standardized output formats [Tang et al. 2024].

Our study investigates how to segment and extract information from a single QFF document while preserving its structural and symbolic layers. We hypothesize that, in documents with recurring and well-defined patterns, a symbolic rule-based approach may provide greater structural control, effectiveness, and computational efficiency than the evaluated LLM-based configuration. This hypothesis is supported by the relevance of rule-based methods for documents with predictable patterns [Li et al. 2008] and by the need for domain-specific comparisons between traditional and neural segmentation approaches [Jegan and Henrich 2025].

To investigate this hypothesis, we compare a symbolic method based on explicit rules, regular expressions, and document heuristics with an LLM-based method implemented through prompt engineering. The evaluation considers precision, recall, F1-score, processing time, structural preservation, and fidelity to the original content. The main contributions are: (i) a symbolic pipeline for segmenting question blocks and speech turns; (ii) a controlled comparison between symbolic and LLM-based extraction; and (iii) an analysis of the errors and structural variations produced by the evaluated approaches.

The remainder of this article is organized as follows. Section 2 presents the related work, Section 3 describes the methodology, Section 4 presents the experiments, Section 5 reports the results, and Section 6 discusses the main findings. Section 7 presents the final considerations and future works.

2. Related Work

Text segmentation and Information Extraction (IE) support the Transformation of unstructured or semi-structured documents into representations suitable for curation, indexing, linguistic analysis, and database construction [Ghinassi et al. 2024]. LLM-based methods have also been applied to structured extraction tasks; for example, [Gazzola et al. 2025] evaluate an LLM for product attribute value extraction against a reference dataset.

However, LLM-based approaches may present difficulties in tasks requiring strict schema adherence and exact preservation of the original structure, including inconsistent responses, output-format variation, and incomplete compliance with predefined schemas

[Tang et al. 2024]. Such limitations are especially relevant when identifiers, speaker attribution, speech-turn boundaries, textual content, and phonetic symbols must be preserved without omission or reorganization.

Symbolic methods based on explicit rules, structural patterns, and regular expressions remain relevant for documents with recurring organization [Li et al. 2008]. Domain knowledge can also support segmentation and normalization [Godoi et al. 2025]. In Portuguese IE research, [Queiroz et al. 2023] describe the transformation of automatically generated extractions into a manually validated corpus, reinforcing the importance of annotation criteria, expert revision, and reliable reference data.

Comparisons between traditional and LLM-based methods emphasize the need for domain-specific evaluation [Jegan and Henrich 2025], while [Xu et al. 2024] discuss the capabilities and limitations of generative approaches to IE. Nevertheless, limited evidence exists on the comparative performance of symbolic and LLM-based methods in linguistic questionnaires that combine question identifiers, graphematic text, phonetic transcriptions, and speech-turn markers under strict structural and symbolic fidelity requirements. Our study addresses this gap by evaluating both approaches on the same Phonetic-Phonological Questionnaire, considering precision, recall, F1-score, processing time, structural preservation, and fidelity to the extracted content.

The next section describes our methodology, including the corpus, preprocessing procedures, extraction strategies, and evaluation protocol.

3. Methodology

Our methodology is experimental and comparative, focusing on the automation of text segmentation and information extraction in a transcribed questionnaire from ALiB Project. In this work, the data are treated as heterogeneous because they bring together graphematic text, phonetic transcriptions, question markings, responses, and speech turns within the same document structure. The methodological objective is to evaluate how different approaches deal with this structure, considering the preservation of linguistic information and the organization of data in a structured format.

The process was divided into stages. Initially, the corpus was analyzed in order to identify the organization of the document, the question markers, the speech turns, and the annotation layers present in the transcription. Next, the document underwent a preprocessing stage aimed at standardizing the file format, preserving the textual content of the phonetic segments, and preparing the data for automatic processing.

Two approaches were implemented and compared. The symbolic approach uses explicit rules, regular expressions, and document heuristics to identify questions, responses, speech turns, and phonetic segments. The LLM-based neural approach uses prompt engineering, with instructions aimed at segmenting and extracting the same information. In both cases, the outputs were organized into structured files, such as CSV, containing fields such as question identification, speech turn, and extracted textual content. Figure 1 presents an overview of the two approaches evaluated.

The comparative evaluation considered effectiveness metrics, such as precision, recall, and F1-score, as well as aspects of computational efficiency, such as processing time and resource usage. The outputs produced by both approaches were compared

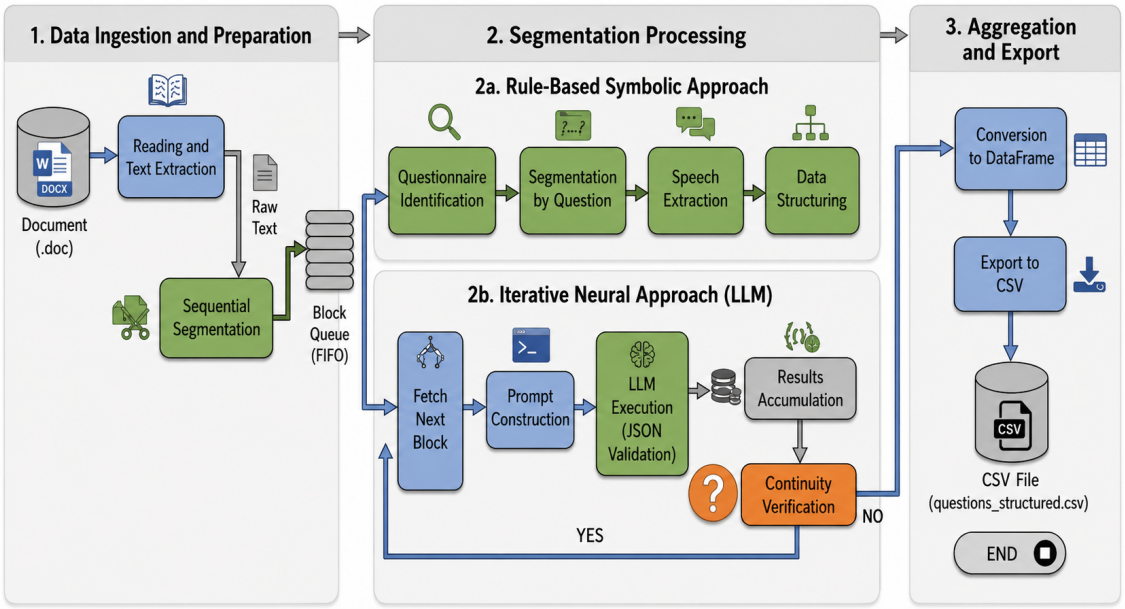


Figure 1. Overview of the symbolic and LLM-based neural approaches.

against a manually constructed gold standard derived from the original questionnaire transcription. This comparison made it possible to evaluate the ability of each method to preserve the document structure, the fields of interest, and the extracted phonetic segments.

3.1. Corpus and Preprocessing

The corpus used in our study consists of a transcribed Phonetic-Phonological Questionnaire (QFF) from Atlas Linguístico do Brasil Project (ALiB) [Cardoso et al. 2014]. The transcription brings together graphematic and phonetic information, as well as question, response, and speech-turn marks. Phonetic forms are represented by symbols from the International Phonetic Alphabet (IPA) [International Phonetic Association 1999].

The initial stage consisted of analyzing the structure of the documents and questionnaires with the objective of understanding the organization of the questions, responses, speech turns, and annotation layers. This process made it possible to define the units to be segmented and the information fields to be extracted, while maintaining alignment with the linguistic characteristics of the corpus.

The handling of phonetic transcriptions required specific care, since the correct visualization of IPA symbols depends on the compatibility among font, encoding, and file format. During data preprocessing, the documents were standardized into a textual format compatible with automatic processing, reducing inconsistencies and textually preserving the phonetic segments.

This task was not necessarily related to loss of information during textual extraction, but to the absence or non-automatic application of appropriate fonts for visualizing IPA symbols. To enable inspection of the results, it was necessary to install the SIL-DoulosIPA and Doulos SIL fonts. Even so, the software used to view the CSV files did not automatically apply these fonts to the phonetic excerpts.

For this reason, an additional manual verification stage was necessary. After the

CSV files were generated, the phonetic segments delimited by brackets were identified and the appropriate font to allow the correct visualization of IPA symbols was applied manually. This stage was validated by a linguist and proved essential for distinguishing real extraction errors from merely visual problems.

3.2. Symbolic Method

The symbolic method performs text segmentation and information extraction deterministically, based on explicit rules defined by regular expressions.

The heuristics used were defined based on the document patterns observed in the questionnaires. These rules make it possible to segment the document into questions and then identify the speech turns associated with each of them.

A) Heuristic 1: Numerical identifiers.

The first heuristic consists of identifying the numerical markers of the questions, represented in the format "(001)", "(002)", "(003)", and so on. These identifiers are used to delimit the beginning of each question and separate the document into structured blocks.

B) Heuristic 2: Speech-turn markers.

The second heuristic consists of identifying speech-turn markers, such as `INF`, `INQ`, `CIR`, and `CIRC`. These markers indicate alternation between the participants in the interaction and make it possible to associate each textual excerpt with its respective turn. When a line does not present a new marker, its content is incorporated into the previous turn, preserving speech segments distributed over more than one line.

3.3. Neural Method

Unlike the symbolic method, this second approach uses a neural method to automatically identify patterns in the transcribed texts. From this, this method categorizes the questions and answers, including the phonemes of the International Phonetic Alphabet (IPA).

The method is based on a pipeline in which the raw text from the file is extracted and divided into smaller context windows, which are used to compose the prompt, where the model receives explicit instructions on how to recognize the structure of the document and what output format it should generate. Instead of relying on previously defined rules, the model infers the patterns directly from the textual context provided.

Different context windows were tested with the objective of evaluating the impact of the size of the excerpt sent to the model both on the precision of the results and on the processing time. This variation made it possible to observe the balance between performance and computational cost.

3.4. Gold Standard Construction and Validation

To evaluate the automated extraction methods, a reference gold standard was manually constructed based on the target Phonetic-Phonological Questionnaire (QFF). To ensure high reliability and address the complexity of linguistic data, the construction and validation process involved three researchers and was carried out in three distinct phases: primary extraction, peer review, and expert consensus.

In the first phase, a primary annotator manually processed the assessment document, systematically extracting speech turns, question identifiers, and responses containing phonetic transcriptions into predefined structured fields. Subsequently, a second researcher reviewed the generated dataset to identify potential omissions or formatting inconsistencies, particularly regarding the preservation of IPA phonetic symbols.

Finally, the validation phase concluded with a consensus meeting involving a linguist. As the task required structural and symbolic fidelity, a consensus-based annotation methodology was adopted rather than parallel blind annotation. All structural alignments and data representations were subject to collective deliberation.

4. Experiments

The experiments were conducted with the objective of comparing the symbolic and neural approaches in text segmentation and extraction of speech-turn extraction from a transcribed QFF document from the ALiB Project. As a case study, the Phonetic-Phonological Questionnaire (QFF), referring to the Salvador 093-02 inquiry, was selected because it contains graphematic and phonetic transcriptions relevant to the proposed evaluation. The comparison considers effectiveness metrics, such as accuracy, precision, recall, and F1-score, as well as aspects of computational efficiency, such as processing time and resource usage. For validation of the results, the outputs produced by the methods were compared with a Gold Standard manually constructed from the original transcriptions.

The evaluation was guided by the following research question:

Q1: Which approach presents the best relationship between effectiveness and computational efficiency in text segmentation and information extraction in a transcribed questionnaire?

4.1. Evaluation of the Segmentation Strategy

In the symbolic approach, the experiments were conducted by applying the heuristics described in Subsection 3.2.1 to the document selected for evaluation. Segmentation was performed deterministically, using numerical identifiers to delimit the questions and speech-turn markers to associate each textual excerpt with its respective participant.

The generated output was structured in CSV format and manually compared with the original document in order to verify the fidelity of the segmentation, the integrity of the extracted content, and the textual preservation of the IPA phonetic transcriptions.

During the initial experiments with the LLM-based neural approach, two strategies for sending the data to the model were compared: processing the complete document and segmentation by context windows. Sending the entire document resulted in a longer response time and more inconsistencies in the extraction, indicating the need to control the volume of information provided in each request.

For this purpose, segmentation by context windows was adopted, in which the file is divided into blocks of fifteen lines of raw text, and this context is injected into the prompt sent to the local model. Since the questionnaire contains 159 questions, the impact of the size of the batches sent to the model was also evaluated. Batches with 10, 15, 25, 50, and 100 context windows were tested, considering processing time, incomplete

or inconsistent responses, and execution stability. Batches greater than 15 questions increased inconsistencies and response variability, while smaller batches increased the total execution time due to the larger number of requests. Thus, the configuration with 15 windows per batch was adopted because it presented the best balance between efficiency and extraction quality.

In order to maintain the privacy of the raw data, the tests using the LLM were performed locally without any exposure to the internet environment through Ollama v0.18.2 [Ollama 2024] with DeepSeek R1 model (8B parameters) [Ollama Library 2024] for identifying, extracting, and categorizing the questions and answers in the transcription documents.

4.2. Experimental Setup

The experiments were carried out on a document from the Phonetic-Phonological Questionnaire (QFF) of the ALiB Project, used as a case study because it contains graphematic and phonetic transcriptions relevant to the segmentation task. The original document, made available in `.doc` format, has 15 pages and contains 159 questions.

After the segmentation of the speech turns, the extractions resulted in CSV files with 211 rows, distributed across three main columns: question identification, speech turn, and textual content. The symbolic and LLM-based neural approaches were applied to the same document, allowing the comparison of the outputs produced under equivalent conditions.

For validation purposes, the outputs generated by the approaches were manually compared with the original document. The comparison aimed to verify the fidelity of the segmentation in relation to the original document, considering the integrity of the textual content and the textual preservation of the IPA phonetic transcriptions.

5. Results

The evaluation was carried out using a document from the Phonetic-Phonological Questionnaire (QFF) from the ALiB Project, composed of 15 pages and 159 questions. After the segmentation of the speech turns, the structured outputs resulted in CSV files with 211 rows, containing the question identification, the speech turn, and the extracted textual content. The outputs of the symbolic and LLM-based neural approaches were compared with a *gold standard* manually constructed from the original transcriptions.

Table 1 presents the execution times observed for the manual process, for the symbolic approach, and for the LLM-based neural approach. In the case of the LLM-based approach, the configuration of batches with 15 questions per request was considered.

Table 1. Execution time observed in the evaluated processes.

Process	Evaluated document	Execution time
Manual	1 QFF	2 h 38 min
Symbolic	1 QFF	0.3 s
LLM-based neural	1 QFF	44 min 13 s

Table 2 presents the quantitative results obtained from the comparison of the automatic outputs with the *gold standard*. The metrics of precision, recall, and F1-score were

considered, in addition to the values of true positives (TP), false positives (FP), and false negatives (FN).

Table 2. Comparison of metrics between the symbolic approach and the LLM-based neural approach.

Metric	Symbolic	LLM-based neural
Precision	98.10%	40.38%
Recall	95.81%	39.07%
F1-score	96.94%	39.72%
TP	206	84
FP	4	124
FN	9	131

The values presented in Table 2 show that the symbolic approach achieved greater correspondence with the *gold standard*. This approach reached an F1-score of 96.94%, while the LLM-based neural approach obtained an F1-score of 39.72%. The difference is also observed in the FP and FN values, indicating that the symbolic approach produced fewer incorrect extractions and omissions in relation to the reference material.

Regarding execution time, Table 1 indicates that the symbolic approach presented lower computational cost. While the manual process required 2 hours and 38 minutes, the symbolic approach processed the document in 0.3 seconds. The LLM-based neural approach, in turn, required 44 minutes and 13 seconds in the evaluated configuration.

The error analysis indicated that the LLM-based neural approach presented greater variability in the organization of the output. In some cases, the model altered the expected structure, such as by dividing the same speech segment into more lines than necessary, impairing the exact correspondence with the original document. The symbolic approach, on the other hand, presented greater structural stability, as it operates based on explicit rules applied to the observed document patterns.

Figure 2 presents examples of the behaviors observed in the outputs of the evaluated approaches. The examples were selected from the manual comparison between the original document, the output of the symbolic approach, and the output of the LLM-based neural approach. The field names *número*, *autor*, and *texto* were preserved in Portuguese because they correspond to the original fields produced by the information extraction process. In this structure, *número* indicates the question number, *autor* identifies the speaker or source, and *texto* contains the extracted textual content.

The qualitative analysis showed that the symbolic approach maintained greater adherence to the original structure, especially in the delimitation of speech turns and in the textual preservation of the phonetic segments. The LLM-based neural approach, although it preserved part of the general content, demonstrated greater variation in the organization of the output, including cases of improper division of speech segments and alteration of the expected structure. These behaviors help explain the differences observed in the quantitative metrics.

Document type	numero	autor	texto
Original Document	1	INF.	- Eh, barraco, casas ['kazɐʃ], apartamentos.
Symbolic	1	INF	- Eh, barraco, casas ['kazɐʃ], apartamentos.
Neural	1	INQ	INF. - Eh, barraco, casas ['kazɐʃ], apartamentos.

Figure 2. Qualitative examples comparing the original document and the outputs of the evaluated approaches.

6. Discussion

Our results, composed of a document from the Phonetic-Phonological Questionnaire (QFF) with 159 questions and 211 rows after the segmentation of speech turns, indicate significant differences between our approaches. Our symbolic approach achieved an F1-score of 96.94%, with 206 true positives, 4 false positives, and 9 false negatives. In comparison, the LLM-based neural approach obtained an F1-score of 39.72%, with 84 true positives, 124 false positives, and 131 false negatives. These values show greater correspondence of the symbolic approach with the *gold standard*, especially due to the lower occurrence of incorrect extractions and omissions.

This difference can be explained by the recurring structure of the analyzed transcribed questionnaires. The presence of numerical identifiers and explicit speech-turn markers favored the application of symbolic heuristics, allowing greater structural control in the segmentation and extraction process. Thus, in documents with well-defined patterns, the predictability of the format contributed directly to the precision, reproducibility, and stability of the symbolic approach.

The LLM-based neural approach presented greater output variability and lower correspondence with the original document. Since this task requires preservation of identifiers, speaker attribution, speech-turn boundaries, textual content, and phonetic symbols, semantic equivalence alone is insufficient. In the evaluated DeepSeek-R1 8B configuration, textual reorganization and improper speech-segment splitting were therefore considered extraction errors [Tang et al. 2024].

Regarding computational efficiency, the difference was also significant. While manual extraction required 2 hours and 38 minutes, the symbolic approach processed the document in 0.3 seconds. The LLM-based neural approach, in the evaluated configuration, required 44 minutes and 13 seconds. These results indicate that, for the analyzed document, the symbolic approach presented the best relationship between effectiveness and computational cost.

These findings reinforce that the choice of extraction strategy should consider the document structure, the required level of fidelity, and the available computational resources. For the analyzed QFF document, the symbolic approach was more suitable because its recurring numerical identifiers and speech-turn markers favored deterministic processing. Although LLM-based methods may be useful when greater contextual inference is required, the evaluated DeepSeek-R1 8B configuration showed lower stability in

a task that required strict structural and symbolic fidelity.

7. Final Considerations and Future Work

This study investigated the feasibility of automating segmentation and information extraction in a transcribed QFF document from the ALiB Project. The comparison between the symbolic approach and the LLM-based neural approach indicated that, in highly structured documents containing phonetic transcriptions, the symbolic approach presents a better balance between precision, computational efficiency, and reproducibility. These results suggest that symbolic methods remain relevant in scenarios in which document patterns are recurrent and structural fidelity is a determining aspect for extraction quality.

As future work, we intend to: (i) Expand the experimental corpus to include multiple ALiB questionnaires from different regions, informants, and document formats, in order to assess the external validity and robustness of the evaluated approaches and support the population of the ALiBWeb relational database; (ii) Include other LLMs and model configurations to assess whether the observed behavior is particular from DeepSeek-R1 8B; (iii) Improve the linguistic analysis based on the structured data, including morphological and dialectological studies; and (iv) Investigate the application of vector representations and phonetic embeddings for more advanced comparative analyses.

Limitations

This study focused primarily on structural and phonetic aspects of the documents, without exploring deeper semantic analyses. Our evaluation was conducted on a single QFF document from the ALiB Project, which limits the generalizability and external validity of the results and does not fully capture the variability of the complete ALiB dataset, including questionnaires from different regions, involving different informants, and available in different document formats. Although the neural approach based on Large Language Models (LLMs) was explored, in this study only one model, DeepSeek-R1 8B, was evaluated. Therefore, the results should not be generalized to LLMs in general, since different models, prompts, and configurations may produce different outcomes. This variability may affect reproducibility and requires careful validation, particularly in structured linguistic data extraction tasks. In addition, the symbolic approach relies on predefined rules and patterns, which may require manual adjustments when applied to documents with different structures or formatting inconsistencies.

Ethics Statement

This work is based on the processing of linguistic data derived from a transcribed QFF document from the ALiB Project. The data contain sensitive personal information that was properly anonymized. All materials were handled in accordance with ethical research standards, with careful attention to data integrity, transparency, reproducibility, and the prevention of harm to individuals or groups. The contact with Internet was prevent during LLM's experiments.

Use of Generative AI Tools

ChatGPT, Gemini, and Copilot were used as auxiliary tools for grammatical revision, text improvement, organization of ideas, English translation, the generation and refinement of visual materials, exploratory searches, the organization of academic references,

and support in the development and revision of Python scripts. All AI-assisted outputs were carefully reviewed, edited, tested, and validated by the authors, who remain fully responsible for the content and conclusions of the final manuscript.

Acknowledgments

This work is supported by INCT TILD-IAR (grant 408490/2024-1), FAPESB CCE022/2023 and FAPESB BOL0129/2025.

References

Cabral, B., Claro, D., and Souza, M. (2024). Exploring open information extraction for Portuguese using large language models. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 127–136, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.

Candido Junior, A., Casanova, E., Soares, A. d. S., Oliveira, F. S. d., Oliveira, L., Fernandes Junior, R. C., Silva, D. P. P. d., Fayet, F. G., Carlotto, B. B., Gris, L. R. S., and Aluísio, S. M. (2023). CORAA ASR: a large corpus of spontaneous and prepared speech manually validated for speech recognition in Brazilian Portuguese. *Language Resources and Evaluation*, 57(3):1139–1171.

Cardoso, S. A. M. d. S., Mota, J. A., Aguilera, V. d. A., Aragão, M. d. S. S. d., Isquerdo, A. N., Razky, A., Margotti, F. W., and Altenhofen, C. V. (2014). *Atlas Linguístico do Brasil*, volume 1. EDUEL, Londrina.

Gazzola, M., Souto, H. G., Silva, S., Peixoto, J. S., Siqueira, F., Morais, A. L. P. d., and Gomes, C. (2025). AI-PAVE-Br: Leveraging large language models for enhanced product attribute value extraction through a golden set approach. In *Proceedings of the Symposium in Information and Human Language Technology (STIL)*.

Ghinassi, I., Wang, L., Newell, C., and Purver, M. (2024). Recent trends in linear text segmentation: A survey. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3084–3095, Miami, Florida, USA. Association for Computational Linguistics.

Godoi, G. A. d., Rivolli, A., Freire, D. L., Pereira, F. S. F., Ventura, N. R., Almeida, A. M. G. d., Garcia, L. P. F., Dias, M. d. S., and Carvalho, A. C. P. d. L. F. d. (2025). Assessing rule-based document segmentation and word normalization for legal ruling classification. In *Conference on Digital Government Research*, volume 26.

International Phonetic Association (1999). *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press, Cambridge.

Jegan, R. and Henrich, A. (2025). Contrasting traditional models and LLMs: An evaluation based on text segmentation. In Wartena, C. and Heid, U., editors, *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*, pages 274–281, Hannover, Germany. HsH Applied Academics.

Krassovitskiy, A., Mussabayev, R., and Yakunin, K. (2025). Llm-enhanced semantic text segmentation. *Applied Sciences*, 15(19).

Li, Y., Krishnamurthy, R., Raghavan, S., Vaithyanathan, S., and Jagadish, H. V. (2008). Regular expression learning for information extraction. In Lapata, M. and Ng, H. T., editors, *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 21–30, Honolulu, Hawaii. Association for Computational Linguistics.

Lu, Y., Liu, Q., Dai, D., Xiao, X., Lin, H., Han, X., Sun, L., and Wu, H. (2022). Unified structure generation for universal information extraction. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5755–5772, Dublin, Ireland. Association for Computational Linguistics.

Ollama (2024). Ollama: Run large language models locally. `https://ollama.com/`. Accessed: 25 fev. 2026.

Ollama Library (2024). Deepseek-r1 8b. `https://ollama.com/library/deepseek-r1:8b`. Large Language Model. Acesso em: 25 fev. 2026.

Queiroz, B., Cavalcante, R., and Claro, D. (2023). Desafios da tarefa de extração de informação aberta: uma abordagem metodológica de um corpus automatizado até o corpus manual. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 388–392, Porto Alegre, RS, Brasil. SBC.

Tang, X., Zong, Y., Phang, J., Zhao, Y., Zhou, W., Cohan, A., and Gerstein, M. (2024). Struc-bench: Are large language models good at generating complex structured tabular data? In Duh, K., Gomez, H., and Bethard, S., editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 12–34, Mexico City, Mexico. Association for Computational Linguistics.

Xu, D., Chen, W., Peng, W., Zhang, C., Xu, T., Zhao, X., Wu, X., Zheng, Y., Wang, Y., and Chen, E. (2024). Large language models for generative information extraction: a survey. *Frontiers of Computer Science*, 18(6):186357.

Exploring Hybrid Pre-training for Automatic Essay Scoring

Miguel M. Carpi[1ǂ], **Igor C. Silveira**[1ǂ], **Denis D. Mauá**[1], **Marcelo Finger**[1]

[1]Department of Computer Science
Institute of Mathematics, Statistics and Computer Science (IME)
University of São Paulo (USP) – São Paulo, SP – Brazil

{miguel,igorcs,ddm,mfinger}@ime.usp.br

Abstract. *Alternatives to Large Language Models have been proposed to develop smaller models that require substantially less training data. In this paper, we propose a monolingual (Portuguese) Hybrid Transformer model trained with only 260M words, whose size is comparable to that of small Encoder-based models. After pre-training, we fine-tune our model and compare it against 11 existing models on the AES-ENEM dataset — an Automatic Essay Scoring benchmark in which models are required to evaluate five distinct textual dimensions. Our experiments demonstrate that our model is always competitive with the (bigger) best available model, despite its smaller scale.*

1. Introduction

Since the introduction of BERT [Devlin et al. 2019], Large Language Models (LLMs) based on the Transformer architecture [Vaswani et al. 2017] have been at the center of Natural Language Processing for most, if not all, languages. This applies to Portuguese, which has at least one (monolingual) model from each of the Transformer families: Encoder [Souza et al. 2020, Rodrigues et al. 2023, Zago and Pedotti 2024], Decoder [Pires et al. 2023, Corrêa et al. 2025a], and Encoder-Decoder [Carmo et al. 2020].

Such models for Portuguese are influenced by (at least) two different sources. The first source involves architectural innovations, which mostly come from English models, such as the changes from BERT to RoBERTa [Liu et al. 2019] and DeBERTa [He et al. 2020]. The second source is the release of Portuguese corpora, such as the GigaVerbo [Corrêa et al. 2025b] and *Corpus* Carolina [Finger et al. 2025]. In general, developers of models for Portuguese tend to follow the neural scaling law, where bigger models — requiring more parameters, data and training time — are assumed to be better.

Although recent advances have largely favored increasingly larger models, smaller language models remain an attractive alternative for specialized tasks or when computational costs are constrained, and they have been argued to represent an important direction for the future of AI systems [Belcak et al. 2025]. Reducing computational costs — memory usage, inference time, and training time — is not a new problem and predates the LLM era [Menghani 2023]. Therefore, a vast literature has been devoted to efficient pre-training and fine-tuning using different techniques, such as knowledge distillation [Hinton et al. 2015], low rank adaptation [Hu et al. 2021], and quantization [Lee et al. 2024]. These and related techniques have been successful in creating small BERT models like nanoBERT [Maity et al. 2024] and Tiny BERT [Jiao et al. 2020].

ǂ Equal contribution

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 62–77
October 19-22, 2026 ©2026 Association for Computational Linguistics

Other approaches showed that BERT-like models require less pre-training data. For example, recent work showed that 100 million words can yield state-of-the-art performance on some NLP benchmarks, although presenting some limitations in more complex tasks [Micheli et al. 2020, Zhang et al. 2021, Samuel et al. 2023]. Moreover, in-context learning was demonstrated in smaller decoders and in encoder-only models [Samuel 2024, Schick and Schütze 2021]. Finally, there is ongoing research in unifying encoder and decoder models in a single architecture to exploit the advantages of each other [Charpentier and Samuel 2024, Yu et al. 2024].

In this work, we propose novel Hybrid Transformer-based models extending previous works [Charpentier and Samuel 2024, Yu et al. 2024, Carpi and Finger 2026]. Our model is approximately the size of a traditional Encoder-only model (150M parameters) but is trained on only 260 million words and has a wider context window than similar Encoders (1024 vs. 512). In order to train this model, a masked-causal hybrid training is employed: the training regime alternates between batches of Masked Language Modeling (Encoder pre-training, through a bidirectional attention mask) and Causal Language Modeling (Decoder pre-training, through a causal attention mask). This creates a model that is capable of behaving like both an Encoder and a Decoder.

To assess the effectiveness of the proposed pre-trained model, we fine-tune and evaluate it on the AES-ENEM dataset [Silveira et al. 2024]. This Automatic Essay Scoring (AES) dataset comprises argumentative essays written for a mock Brazilian national university entrance exam, the ENEM. We selected this dataset for two main reasons. First, the task requires the model to adapt to five different aspects of written language, posing a significant and comprehensive challenge. Second, 15 different LLMs were benchmarked in this dataset [Barbosa et al. 2025], allowing for direct comparison against many models with only one experiment. When comparing our best performance against the benchmarked results, we see that: I) in one textual dimension (argumentative quality) it outperformed the others — but it was still statistically similar to the previous best; II) in another dimension (persuasiveness) it had an inferior performance, but the confidence interval still overlaps with the best models; III) in the remaining three dimensions, our model exhibited a performance similar to the best model, losing by only a small margin — remaining inside the confidence interval.

As a by-product of our different combinations of training regimes, we can create a spectrum that goes from Decoders to Encoders. Thus, we are capable of evaluating the influence of the training regime on different traits, which has never been assessed before. Additionally, we contribute to AES with a new baseline performance: the performance of a Transformer model without pre-training.

To summarize, our main contributions are:

- presenting a family of Hybrid models trained with limited data;[1]
- showing that our best models are statistically comparable to the SOTA;
- filling the gap in AES regarding the importance of the pre-training regime;
- presenting a Transformer baseline performance for the AES-ENEM benchmark.

The remainder of the document is divided as follows: related work, presentation of our hybrid models, methodology, results, and conclusion.

[1] Available at: `https://github.com/mmcarpi/flexqwen`

Table 1. Encoder-based and Decoder (Small LLMs) benchmarked in [Barbosa et al. 2025] together with their context and model size.

Model Name	Architecture	Context	#Params
BERTimbau Base	Encoder	512	109M
BERTuguês	Encoder	512	110M
mBERT	Encoder	512	178M
BERTimbau Large	Encoder	512	334M
Albertina	Encoder	512	1.5B
Phi3	Decoder	4k	2B
Tucano	Decoder	4k	2.4B
Llama3	Decoder	128k	8B
Phi4	Decoder	16k	14.7B

2. Related Work

In this section, we discuss the two topics related to our research. We started by situating our model in the LLM landscape. Then, we review models for AES in Portuguese.

Traditionally, Transformer-based models can be categorized according to their architecture — Encoder, Decoder, or Encoder-Decoder — and pre-training regime — Masked Language Modeling or Causal Language Modeling. Pure-Encoder models have an Encoder architecture and are pre-trained using Masked Language Modeling. Such models usually excel at language understanding and are used for tasks such as text classification. One advantage of this group is that they are relatively small, the most popular ones used for Portuguese — BERTimbau [Souza et al. 2020], BERTuguês [Zago and Pedotti 2024], and Albertina [Rodrigues et al. 2023] — range from 109M to 1.5B parameters. Their disadvantage is their limited context window of 512 tokens.

Pure-Decoder models are those that use Decoder architecture and are pre-trained using Causal Language Modeling. This class can be further divided. The first division is API Decoders, as they are usually accessed through APIs due to their huge size and are often proprietary, making them unable to be fine-tuned on everyday desktops. Examples of these models are Sabiá3 [Pires et al. 2023], GPTo1 [OpenAI et al. 2024], and DeepSeek-R1 [DeepSeek-AI et al. 2025]. The second group is often called Small Large Language Models because they are smaller than the previous group, but still orders of magnitude bigger than Encoder models. Examples of these models are some variants of Llama3 [Touvron et al. 2023], Qwen3 [Yang et al. 2025], Phi [Abdin et al. 2024a, Abdin et al. 2024b], and Tucano [Corrêa et al. 2025a]. These models can be fine-tuned, but they usually require high-end equipment that is not commonly available.

Finally, Encoder-Decoder models have the architecture of the same name and are pre-trained using both of the pre-training regimes. Although they are available in Portuguese through the PTT5 models [Carmo et al. 2020, Piau et al. 2024], they are not as widely adopted as the previous groups.

Our Hybrid model differs from all the previous ones as it uses a single architecture and is trained with both pre-training regimes, thus being able to behave like an encoder- or decoder-only model. Additionally, it has a smaller size, as an Encoder, it is trained on a small portion of data and has an increased context of 1024 tokens that can be further

extended thanks to the employment of positional encoding through RoPE [Su et al. 2024].

AES of ENEM-like essays has a tradition of employing all the stages of Natural Language Processing technologies. In the early days, approaches were based on word occurrence [Bazelato and Amorim 2013] and features that were proxies for textual quality [Amorim and Veloso 2017]. In the neural network era, feature extraction was delegated through fine-tuning, first to fixed dense representations and LSTMs [Fonseca et al. 2018, Marinho et al. 2022], then to Encoder models [Silveira et al. 2024], and later to Small LLMs [Silveira et al. 2025a]. Finally, modern conversational agents with reasoning were employed through zero-shot prompting [Barbosa et al. 2025]. Noteworthy, due to their smaller size, Encoder Transformers are widely employed for Portuguese AES, either directly [Ribeiro et al. 2024, de Sousa et al. 2024, Rossman et al. 2026] or indirectly through majority voting [de Lima et al. 2024], Two-Stage Learning [Silveira et al. 2025b], combined representations [Matos and Feltrim 2026], or neuro-symbolic methods [Silveira and Mauá 2026].

Several models, ranging from feature-based to modern reasoning-based API Decoders, have been benchmarked [Barbosa et al. 2025] on the same ENEM-like AES dataset [Silveira et al. 2024]. This allows any new model to be directly compared to them. The models that are comparable to ours and that were benchmarked in the aforementioned work are summarized in Table 1. According to the authors of the benchmarking, there was no model that was consistently better than the others, highlighting that each model may excel in a different dimension of language. Alternatively, the impact of different representations used by AES systems was investigated in [Chalegre et al. 2026]. The authors of the previous study concluded that although contextual representations given by Encoders have advantages in mechanical traits, TF-IDF-based representations could be competitive in thematic traits, which are usually dominated by Decoders in the benchmark.

The present work intersects with AES literature in two different points. The first is by evaluating a novel family of Hybrid Transformer models, which have the same number of parameters as Encoder models but a larger context, in the same benchmark. The second is by evaluating the same LLM with different pre-training regimes, which is still not covered in the literature.

3. Hybrid Model

For the model configuration, a model is instantiated from scratch based on the architectural specifications of Qwen 3 [Yang et al. 2025], rather than from a pre-trained checkpoint. To increase neural network expressivity, we additionally incorporate layer dropout and layer weighting [Charpentier and Samuel 2023]. To support our hybrid pre-training objective, the model contains a total of 152,981,916 trainable parameters, featuring a vocabulary size of 64,000, a hidden dimension of 768, an intermediate feed-forward dimension of 3,072, 12 transformer blocks, 12 attention heads with a head dimension of 64, and 4 key-value groups. Although we could have used the original Qwen 3 tokenizer, or another existing multilingual or Portuguese-specific tokenizer, we trained a custom byte-pair encoding tokenizer [Sennrich et al. 2016] with a vocabulary size of 64,000 on a reserved split of our *corpus* (described below) to ensure tighter alignment between the vocabulary and our pre-training data. Before chunking, text is normalized using Unicode Normalization Form C so that accented characters (e.g., *ã*, *é*) map to unified code points.

Table 2. Model performance on test set by hybrid strategy and objective function: Causal Language Modeling (CLM) and Masked Language Modeling (MLM). Models were evaluated in sequences of 1024 tokens. The corruption for MLM was the same as the one used for training (30%).

Objective	Model	Perplexity	Loss
CLM	causal	**15.34**	**2.73**
	constant	17.94	2.88
	linear	20.42	3.01
	cosine	20.55	3.02
	shift	21.52	3.06
MLM	masked	17.23	2.84
	constant	17.90	2.88
	linear	17.29	2.85
	cosine	17.22	2.84
	shift	**17.04**	**2.83**

This study exclusively utilizes the *Corpus* Carolina v2.0 [Finger et al. 2025], a curated Brazilian Portuguese web dataset, for language model pre-training. We implemented a two-stage filtering pipeline following established data-curation practices for pre-training large language models [Rae et al. 2021]. First, we applied typology-aware length thresholds — excluding the top and bottom 5% of texts — and removed low-entropy documents based on unique word ratios, punctuation density, and extreme word lengths, thus removing texts that are too short, or that contains an excessive number of URLs or markup text. Second, to eliminate near-duplicate documents that can lead to memorization and degraded generalization [Lee et al. 2022, Leveling et al. 2024], we applied MinHash-based Jaccard similarity. This pipeline reduced collectively the *corpus* from approximately 2.07 million to 1.78 million documents. From this refined dataset, a final sample of 1.5 million documents, roughly 320 million words, was extracted and partitioned into 260 million words for model training, 50 million for tokenizer development, and 5 million each for validation and testing.

Furthermore, we investigate a family of hybrid Transformer models capable of functioning as either encoder-only or decoder-only architectures purely through attention mask modifications. Building on prior research in hybrid causal-masked pre-training [Charpentier and Samuel 2024, Yu et al. 2024, Carpi and Finger 2026], we evaluate four objective-mixing strategies that dictate the probability of applying a Causal Language Modeling (CLM) versus a Masked Language Modeling (MLM) objective during training: Shift (a sequential transition from CLM to MLM), Constant (a fixed CLM probability throughout training), Linear (a linear reduction of the CLM probability over time), and Cosine (a cosine decay applied to the CLM probability schedule). These strategies were influenced by past works. The Shift strategy originates from [Yu et al. 2024] which studies taking a causal model and further training it on the masked next-token prediction to achieve an encoder model. The Linear strategy is adapted from [Carpi and Finger 2026] which proposes alternating batches of MLM and CLM optimization. The Constant strategy is an adaptation from [Charpentier and Samuel 2024], where instead of optimizing both CLM and MLM objectives in every batch, we choose with some probability p if a particular batch should be used to optimize the CLM or the MLM objective. To the best

Table 3. Variation space for fine-tuning. The models masked, causal and non-pre-trained correspond to our baselines. Note that we do not fine-tune the causal model with a bidirectional mask nor the masked model with a causal mask.

Parameter	Values
Model	masked, shift, constant, linear, causal, cosine, non-pretrained
Input	essay, full
Pooling	last token, average
Pre-training	bidirectional, causal

of our knowledge, while the cosine decay is widely used as a learning rate scheduler, its application to balancing CLM and MLM objective-mixing ratios has not yet appeared in the literature.

We pre-trained six models — a CLM baseline, an MLM baseline, and the four aforementioned hybrid variants — for four epochs with a batch size of 524,288 tokens. Optimization relied on Muon [Keller Jordan et al. 2024] (learning rate 0.02) for 2D matrices and AdamW [Loshchilov and Hutter 2018] (learning rate 3×10^{-3}) for the embedding layer, applying a 0.95 momentum, 0.02 weight decay, zero dropout, and a cosine-decay scheduler with a 5% warmup. Training featured a progressive sequence length warmup (128, 512, and 1024 tokens across the first three epochs), while validation strictly evaluated 1024-token sequences. For masked and hybrid architectures, inputs were corrupted at a 30% rate using span masking (lengths 3 to 10) [Joshi et al. 2020]. Hybrid objective ratios were governed by either maintaining a fixed 0.55 CLM probability (constant strategy) or dynamically decaying the CLM probability from 0.9 to 0.2 over the first 55% of the training steps (cosine, linear, and shift strategies).

The pre-training results in Table 2 show that our hybrid models performed worse than the decoder-only baseline (causal) in the CLM objective, while in the MLM objective the shift strategy beat the encoder-only (masked). Additionally, the constant strategy resulted in a model that performs roughly the same in both CLM and MLM, and not so distant from the baselines. The other hybrid models present a higher loss in the CLM objective in comparison with their loss in MLM, which is expected since they were trained for longer in the MLM objective.

4. Methodology

To fine-tune and assess the models, we employ the same split used in the benchmarking of the other models [Barbosa et al. 2025] — a filtered version of AES-ENEM [Silveira et al. 2024]. In this filtered version, 385 mock essays were evaluated by two experienced graders according to the five traits of the exam: C1) proper usage of Portuguese, C2) thematic coherence, C3) argumentation quality, C4) cohesion, C5) persuasiveness of conclusion. Each trait is evaluated on a scale from 0 to 200, in steps of 40. Noteworthy, each essay appears twice, as it was graded by two graders.

The models are evaluated according to the Quadratic Weighted Kappa (QWK) metric, which is widely used for AES [Doewes et al. 2023]. This metric ranges from 1 (perfect agreement) to -1 (perfect disagreement), where 0 indicates random agreement. The expected random agreement is calculated based on the marginal distributions of the

Table 4. Comparison of model performance using Bidirectional mode for essay-only input. In bold, the highest scoring model inside each trait and each pooling strategy

Mode: Bidirectional												
model	C1	C2	C3	C4	C5	avg.	C1	C2	C3	C4	C5	avg.
Pool. Strategy:	*average*						*last token*					
constant	**.64**	.22	.33	.47	.40	.41	.44	.13	.16	.36	**.47**	.31
cosine	.54	**.28**	.28	.40	.39	.37	**.59**	.16	.26	**.56**	.31	.37
linear	.49	.02	.22	.44	.04	.24	.51	**.23**	.28	.39	.37	.35
masked	.46	-.05	**.36**	.36	.35	.29	.44	.15	**.31**	.40	.28	.31
shift	.52	-.09	.35	.37	**.41**	.31	.50	.13	.25	.45	.44	.35
non-pretrained	.16	.16	.16	**.48**	.03	.19	.22	.16	.25	.36	.11	.22

labels assigned by the annotators being evaluated, and disagreement is penalized quadratically based on the distance between the ordinal labels. The reported performance of the model is the average of QWK with each grader. Finally, we followed the same approach from the benchmark and calculated the 95% confidence interval ($N = 1000$) of each model, so that we could assert whether the differences in performance are significant.

Our models were fine-tuned using the AdamW optimizer (learning rate of 5×10^{-5}, weight decay of 0.05 excluding biases and RMSNorm, and dropout of 0.05) alongside a cosine-decay scheduler with a 10% warmup, training for up to 20 epochs with a batch size of 32 and a four-epoch early stopping patience based on validation QWK scores. Our experimental framework evaluates a *non-pretrained* baseline against several architectural and input variations: pooling strategy (last token versus average representation), training mode (causal versus bidirectional masking), and input strategy (*essay-only* versus a *full* context exceeding the 1024-token pre-training limit to assess length generalization). We add a non-pretrained model to assess the impact of pre-training for the different traits. Table 3 summarizes our variation space.

5. Results

As our models have a large variation space, we structure our analysis first by input strategy. Then, we segment training mode and pooling strategies across all five traits. Finally, we establish the highest-performing configurations and compare our findings with the previous literature.

Starting with essay-only input, we present in Table 4 the results of training with a bidirectional mask. As with previous research, there is no model that consistently outperforms all the others. The only case where dominance occurs is in the average pooling, between the constant and linear models. When comparing the model with the highest average performance against the non-pretrained baseline, we see that pre-training was advantageous. However, its importance changes according to the model — importantly, coherence (C4) in the average pooling was better assessed without pre-training. When comparing the pooling strategies, we can see that averaging the representations caused the models to have a higher variance — their average performances ranges from .24 to .41, while using the last token ranges from .31 to .37.

In Table 5, we present the results when the training regime was causal. Here, there

Table 5. Comparison of model performance using Causal mode for essay-only input. In bold, the highest scoring model inside each trait and each pooling strategy

Mode: Causal												
model	C1	C2	C3	C4	C5	avg.	C1	C2	C3	C4	C5	avg.
Pool. Strategy:	*average*						*last token*					
causal	**.47**	.01	.26	.35	-.01	.21	**.57**	.19	.32	.37	.35	.36
constant	.41	.00	**.27**	**.47**	.08	.24	.55	.21	.22	.41	.44	.36
cosine	.42	.00	.24	.43	.02	.22	.46	**.30**	.29	.47	.40	.38
linear	.37	**.02**	.20	.33	.08	.20	.47	.23	**.42**	.38	.38	.37
shift	**.47**	-.25	.13	.38	**.28**	.20	.46	.18	.12	**.48**	**.49**	.34
non-pretrained	.41	.12	.24	.20	-.02	.19	.11	.13	.15	.45	.17	.20

Table 6. Lower and upper bound of the best hybrid model by trait

trait	lower	upper	QWK	model	classifier head	mode
C1	.54	.74	.64	constant	reduce mean	Bidirectional
C2	.13	.48	.30	cosine	last token	Causal
C3	.24	.56	.42	linear	last token	Causal
C4	.42	.69	.56	cosine	last token	Bidirectional
C5	.31	.63	.49	shift	last token	Causal

are no cases of dominance. We see that the models with average pooling did not benefit from pre-training. The last token pooling, on the other hand, did benefit from it. Here, both pooling strategies yielded models that are very similar on average: they present only .04 points of amplitude in both cases.

When analyzing based on the best performance in each trait, we see an interesting result: the average pooling appears as the best strategy in only one case (C1). The bidirectional mode was the best in two traits, C1 and C4, both previously identified as traits better evaluated by Encoders [Barbosa et al. 2025] — which are bidirectional. Lastly, the pure models (causal model with causal mode and masked model with bidirectional mode) are never the best performing. The best ones are, respectively: constant, cosine, linear, cosine, and shift — all of them are hybrid models. In Table 6, we present the best scoring configurations for each trait, together with their lower and upper bounds of performance according to the bootstrap method.

Next, we investigate the usage of the full context to grade the essays. This step is relevant because C2 and C3 require relating the essay to the prompt and the supporting text given in the context. Ideally, adding the context should not hurt the performance in the other three traits, but it has been noted that this is not the case [Barbosa et al. 2025]. We present the best performances in the first division of Table 7, where we refer to our models as FlexQwen.

In general, we see that the performance either remained the same (C5) or decreased when adding the full context. One explanation for this observation is that the model did not receive input sequences with length longer than 1024 during training. Thus we conducted an experiment whre the model was pre-trained for one more epoch with

Table 7. Performance of best performing models: ours and previously benchmarked.

	C1	C2	C3	C4	C5
Best FlexQwen (essay-only)	.64	.30	.42	.56	.49
Best FlexQwen (full context)	.50	.24	.26	.46	.49
Best Encoder [Barbosa et al. 2025]	.68	.32	.29	.60	.63
Best Small LLM [Barbosa et al. 2025]	.66	.36	.37	.49	.59

sentences of length 2048 and then fine-tuned for C2 and C3, but the results were not significantly higher.

In the second division of Table 7, we present the best model performance in each trait as presented in the benchmark. Importantly, these performances come from the experiments where the models only had access to the essay text. The best Encoder are: a draw between Albertina and BERTimbau-large for C1, BERTimbau-large for C2, BERTuguês for C3, BERTimbau-large for C4, and BERTimbau-base for C5. While the best Decoders are, respectively: Tucano, Llama3, Llama3, Tucano, Phi4.

Three cases emerge when we compare our best essay-only models against the selected models from the benchmark. The first case occurs in C1 and C2: the SOTA is not that far from our model's performance (.06) and is within the confidence interval. The second case occurs in C3 and C4, where our model is better than at least one of the SOTA — when our model has better performance, the other model is within its confidence interval. The last case comprises C5, where the SOTA is the upper bound of our model. With this comparison, we see that our model, created with limited data, is comparable to models with many more parameters and trained for longer time on more data.

As a final comparison, we extend the baseline performances presented in the benchmark. There, the authors used a Linear Regressor and Random Forest over textual features calculated by NILC-Metrix [Leal et al. 2024]. With our work, it is possible to select the best non-pretrained model and present it as the baseline performance of the Transformer architecture. In addition, in the benchmark, the confidence interval of the feature-based models was not presented.[2] We calculate for our models and present the comparison in Table 8.

It is interesting to note that in the mechanical traits (C1 and C4), the hybrid models had the same performance as the feature-based, that use syntax- and grammar-related features. This further supports the hypothesis that Transformer-based models are approximating these features when fine-tuned on this dataset [Silveira et al. 2025a]. For the two main thematic traits (C2 and C3), our models performed worse than the feature-based ones, suggesting that the models were not capable of identifying which features would be important for these cases. Finally, for C5, our model is slightly better, probably because in this trait, some formulations are important — which are better identified through the words rather than proxies — but they also require some world knowledge, that is why it was not able to achieve a good performance. Given that, for the mechanical traits, the non-pretrained model achieves nearly .4 QWK and the pre-trained reach nearly .6, it is

[2]Their confidence intervals are presented at: `https://huggingface.co/datasets/kamel-usp/jbcs2025_experiments_report`

Table 8. Lower and upper bound of best non-pretrained hybrid model by trait and best performance from Feature-based models in the benchmark.

	Hybrid			Best Feature-based		
Trait	**Lower**	**Upper**	**QWK**	**Lower**	**Upper**	**QWK**
C1	.26	.53	.41	—	—	.41
C2	.00	.32	.16	—	—	.32
C3	.05	.44	.25	—	—	.35
C4	.35	.62	.48	—	—	.48
C5	-.01	.35	.17	—	—	.12

possible that better models can be created for these traits by using fewer Transformer blocks or by pre-training for less time.

6. Conclusion

In this work, we presented contributions in two different areas. The first, and main, area is the proposal of a family of Hybrid Transformer-based models trained with a relatively small amount of data. Then, through Automatic Essay Scoring, we validated that this model is competitive with much larger models, such as Albertina and Phi4, in evaluating five different textual dimensions — ranging from proper usage of Portuguese to argumentative quality. This suggests that it is possible to create better models for Portuguese without relying on increasing their scale.

The second area to which we contribute is Automatic Essay Scoring. Previous baseline performances were based on feature-based Linear Regressors and Random Forest. Our non-pretrained model fills the role of a Transformers-based baseline model, as it presents the performance of this architecture without pre-training. Previous research has shown that mechanical traits (C1 and C4) were better graded by Encoders. In our experiments, we showed that the bidirectional mode of training — typically used to pre-train Encoders — yielded the best performing models for these traits. Also, in accordance with previous research, thematic competencies were better evaluated by models trained with causal mask. One other contribution of our work was presenting that the best scoring model although following the trend of previous research, was not a pure masked model for mechanical traits nor a pure causal model for the thematic ones. This suggests that combining both pre-training regimes may lead to better models. Finally, the analysis of our non-pretrained model may suggest that for mechanical traits, it may be advantageous to design models with fewer Transformer blocks, making the model even more manageable.

Limitations

For our experiments, we employed the same dataset split and methods used in [Barbosa et al. 2025] so that we could compare against the previously evaluated models. One downside of this approach is that the bootstrap method with a relatively small dataset causes the confidence intervals to be large, which induces more false positives. In other words, some models may seem like equivalents without actually being so.

The capacity of our Hybrid models was not exhausted. It is possible to further improve its context window by presenting sentences of 2048 tokens or longer during

training. Finally, it is also possible to use it as a Decoder and prompt it to assess the essays. We focused on evaluating it as a traditional Encoder, as it reduces the number of experiments and makes the output simpler.

Ethics Statement

This research utilizes the *Corpus* Carolina v2.0 for pre-training and a previously established, anonymized dataset of mock ENEM essays for fine-tuning and evaluation. All data used in this study is publicly available and properly licensed for research purposes. We recognize that language models can inadvertently encode or amplify linguistic biases present in their training data, potentially disproportionately affecting marginalized dialects or non-standard varieties of Portuguese. While our hybrid model demonstrates competitive performance, we advocate that such automatic essay scoring systems should be seen as research objects. If employed as systems, they should be used as assistive tools for educators rather than autonomous grading authorities, ensuring human oversight in high-stakes educational assessments.

Disclosure of Generative AI Usage

During the preparation of this manuscript, the authors utilized generative AI to assist with English grammar revision and structural proofreading. AI tools were also used for finding bugs in experimental code. All core research, conceptualization, data analysis, and final content decisions were solely performed by the human authors.

Aknowledgements

This work was carried out at the Center for Artificial Intelligence (C4AI-USP), with support by the University of São Paulo, the São Paulo Research Foundation (FAPESP) (grant 2019/07665-4). Marcelo Finger was partly supported by the São Paulo Research Foundation (FAPESP) (grants 2023/00488-5 (SPIRA-BM), 2022/11254-2 (EMU)); and the National Council for Scientific and Technological Development (CNPq) (grant PQ1 302963/2022-7). Denis Deratani Mauá was partly supported by FAPESP (grant 2022/02937-9) and CNPq (grant 305136/2022-4). Igor Silveira was supported by FAPESP (grant 2026/15768-1). This study was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) — Finance Code 001.

References

Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A. A., Bach, N., Bahree, A., Bakhtiari, A., Bao, J., Behl, H., Benhaim, A., Bilenko, M., Bjorck, J., Bubeck, S., , et al. (2024a). Phi-3 technical report: A highly capable language model locally on your phone.

Abdin, M., Aneja, J., Behl, H., Bubeck, S., Eldan, R., Gunasekar, S., Harrison, M., Hewett, R. J., Javaheripi, M., Kauffmann, P., et al. (2024b). Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Amorim, E. C. F. and Veloso, A. (2017). A multi-aspect analysis of automatic essay scoring for Brazilian Portuguese. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 94–102. Association for Computational Linguistics.

Barbosa, A., Silveira, I. C., and Mauá, D. D. (2025). An empirical analysis of large language models for automated cross-prompt essay trait scoring in brazilian portuguese. *Journal of the Brazilian Computer Society*, 31(1):857–870.

Bazelato, B. S. and Amorim, E. C. F. (2013). A bayesian classifier to automatic correction of portuguese essays. In *Conferência Internacional sobre Informática na Educação (TISE)*, volume 18, pages 779–782.

Belcak, P., Heinrich, G., Diao, S., Fu, Y., Dong, X., Muralidharan, S., Lin, Y. C., and Molchanov, P. (2025). Small Language Models are the Future of Agentic AI.

Carmo, D., Piau, M., Campiotti, I., Nogueira, R., and Lotufo, R. (2020). Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. *arXiv preprint arXiv:2008.09144*.

Carpi, M. d. M. and Finger, M. (2026). FlexQwen: Exploring Hybrid Objectives and Text Originality for Portuguese. In Souza, M., de-Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 1079–1084, Salvador, Brazil. Association for Computational Linguistics.

Chalegre, P. C., Machado, V. d. R., and Feltrim, V. D. (2026). Avaliação automática de redações do enem: Uma análise comparativa entre engenharia de características e transformers. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 738–748, Salvador, Brazil. Association for Computational Linguistics.

Charpentier, L. G. G. and Samuel, D. (2023). Not all layers are equally as important: Every Layer Counts BERT. In Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjabe, B., Williams, A., Linzen, T., and Cotterell, R., editors, *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 238–252, Singapore. Association for Computational Linguistics.

Charpentier, L. G. G. and Samuel, D. (2024). GPT or BERT: Why not both? In Hu, M. Y., Mueller, A., Ross, C., Williams, A., Linzen, T., Zhuang, C., Choshen, L., Cotterell, R., Warstadt, A., and Wilcox, E. G., editors, *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 262–283, Miami, FL, USA. Association for Computational Linguistics.

Corrêa, N. K., Sen, A., Falk, S., and Fatimah, S. (2025a). Tucano: Advancing Neural Text Generation for Portuguese. *Patterns*.

Corrêa, N. K., Sen, A., Falk, S., and Fatimah, S. (2025b). Tucano: Advancing neural text generation for Portuguese. *Patterns*, 6(11):101325.

de Lima, T. B., Freitas, E., and Macario, V. (2024). Aesvoting: Automatic essay scoring with bert and voting classifiers. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 2*, pages 6–9.

de Sousa, R. F., Marinho, J. C., Neto, F. A., Anchiêta, R., and Moura, R. S. (2024). PiLN at PROPOR: A BERT-Based Strategy for Grading Narrative Essays. In *Proceedings of*

the 16th International Conference on Computational Processing of Portuguese-Vol. 2, pages 10–13.

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., and et al. (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Doewes, A., Kurdhi, N. A., and Saxena, A. (2023). Evaluating quadratic weighted kappa as the standard performance metric for automated essay scoring. In Feng, M., KÃ¤ser, T., and Talukdar, P., editors, *Proceedings of the 16th International Conference on Educational Data Mining*, pages 103–113, Bengaluru, India. International Educational Data Mining Society.

Finger, M., de Sousa, M. C. P., Namiuti, C., do Monte, V. M., Costa, A. S., Serras, F. R., Sturzeneker, M. L., Carpi, M. d. M., Palma, M. F., and Lachi, G. A. (2025). Building Carolina: Metadata for Provenance and Typology in a Corpus of Contemporary Brazilian Portuguese. *Cadernos de Linguística*, 6(4):e812–e812.

Fonseca, E. R., Medeiros, I., Kamikawachi, D., and Bokan, A. (2018). Automatically grading brazilian student essays. In *Computational Processing of the Portuguese Language. PROPOR 2018.*, pages 170–179.

He, P., Liu, X., Gao, J., and Chen, W. (2020). DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION. In *International Conference on Learning Representations*.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the Knowledge in a Neural Network.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., and Liu, Q. (2020). TinyBERT: Distilling BERT for Natural Language Understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., and Levy, O. (2020). SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein (2024). Muon: An optimizer for hidden layers in neural networks | Keller Jordan blog. https://kellerjordan.github.io/posts/muon/.

Leal, S. E., Duran, M. S., Scarton, C. E., Hartmann, N. S., and Aluísio, S. M. (2024). Nilc-metrix: assessing the complexity of written and spoken language in brazilian portuguese. *Language Resources and Evaluation*, 58(1):73–110.

Lee, C., Jin, J.-g., Cho, Y., and Park, E. (2024). QEFT: Quantization for Efficient Fine-Tuning of LLMs. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Findings*

of the Association for Computational Linguistics: EMNLP 2024, pages 13823–13837, Miami, Florida, USA. Association for Computational Linguistics.

Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., and Carlini, N. (2022). Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445.

Leveling, J., Helmer, L., Stein, B. J., Wegener, D., Sheikh, Z., Fernandes, E., and Abdelwahab, H. (2024). Evaluation of document deduplication algorithms for large text corpora. In *International Conference on Machine Learning, Optimization, and Data Science*, pages 390–404. Springer.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Loshchilov, I. and Hutter, F. (2018). Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Maity, K., Chaulwar, A. T., Vala, V., and Guntur, R. S. (2024). NanoBERT: An Extremely Compact Language Model. In *Proceedings of the 7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD CODS and 29th COMAD)*, CODS-COMAD '24, pages 342–349, New York, NY, USA. Association for Computing Machinery.

Marinho, J. C., Cordeiro, F., Anchiêta, R. T., and Moura, R. S. (2022). Automated essay scoring: An approach based on enem competencies. In *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional*, pages 49–60. SBC.

Matos, G. G. d. and Feltrim, V. D. (2026). Avaliação automática de redações do enem: Um estudo empírico sobre representações linguísticas e contextuais. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 488–497, Salvador, Brazil. Association for Computational Linguistics.

Menghani, G. (2023). Efficient Deep Learning: A Survey on Making Deep Learning Models Smaller, Faster, and Better. *ACM Computing Surveys*, 55(12):1–37.

Micheli, V., d'Hoffschmidt, M., and Fleuret, F. (2020). On the importance of pre-training data volume for compact language models. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7853–7858, Online. Association for Computational Linguistics.

OpenAI, :, Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., and et al. (2024). Openai o1 system card.

Piau, M., Lotufo, R., and Nogueira, R. (2024). ptt5-v2: A closer look at continued pretraining of t5 models for the portuguese language. In *Brazilian Conference on Intelligent Systems*, pages 324–338. Springer.

Pires, R., Abonizio, H., Almeida, T. S., and Nogueira, R. (2023). *Sabiá: Portuguese Large Language Models*, page 226–240. Springer Nature Switzerland.

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., et al. (2021). Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Ribeiro, E., Mamede, N., and Baptista, J. (2024). Exploring the automated scoring of narrative essays in brazilian portuguese using transformer models. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 2*, pages 14–17.

Rodrigues, J., Gomes, L., Silva, J., Branco, A., Santos, R., Cardoso, H. L., and Osório, T. (2023). Advancing neural encoding of portuguese with transformer albertina pt-*.

Rossman, L. N., Silveira, I. C., and Mauá, D. D. (2026). Evaluating automated scoring models on official ENEM essays. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 161–171, Salvador, Brazil. Association for Computational Linguistics.

Samuel, D. (2024). Berts are generative in-context learners. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C., editors, *Advances in Neural Information Processing Systems*, volume 37, pages 2558–2589. Curran Associates, Inc.

Samuel, D., Kutuzov, A., Øvrelid, L., and Velldal, E. (2023). Trained on 100 million words and still in shape: BERT meets British National Corpus. In Vlachos, A. and Augenstein, I., editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1954–1974, Dubrovnik, Croatia. Association for Computational Linguistics.

Schick, T. and Schütze, H. (2021). It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. In Erk, K. and Smith, N. A., editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Silveira, I. C., Barbosa, A., da Costa, D. S. L., and Mauá, D. D. (2025a). Investigating universal adversarial attacks against transformers-based automatic essay scoring systems. In Paes, A. and Verri, F. A. N., editors, *Intelligent Systems*, pages 169–183, Cham. Springer Nature Switzerland.

Silveira, I. C., Barbosa, A., and Mauá, D. D. (2024). A new benchmark for automatic essay scoring in Portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 228–237.

Silveira, I. C. and Mauá, D. D. (2026). Neuro-symbolic approaches for rubric-based automatic essay evaluation of ENEM essays. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 790–799, Salvador, Brazil. Association for Computational Linguistics.

Silveira, I. C., Ribeiro, E., Mamede, N., and Baptista, J. (2025b). Aprendizado por transferência para correçao automática de redaçao. *Linguamática*, 17(2):99–116.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. (2024). RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing*, 568:127063.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. (2025). Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yu, X., Guo, B., Luo, S., Wang, J., Ji, T., and Wu, Y. (2024). AntLM: Bridging Causal and Masked Language Models. https://arxiv.org/abs/2412.03275v1.

Zago, R. M. and Pedotti, L. A. d. S. (2024). Bertugues: A novel bert transformer model pre-trained for brazilian portuguese. *Semina: Ciências Exatas e Tecnológicas*, 45:e50630.

Zhang, Y., Warstadt, A., Li, X., and Bowman, S. R. (2021). When Do You Need Billions of Words of Pretraining Data? In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1112–1125, Online. Association for Computational Linguistics.

ENEM Essay Feedback Using LLM-Augmented Prompts

Flavio Carvalho[1], **Vanessa Soares**[1], **Eduardo Bezerra**[1], **Gustavo Guedes**[1]

[1]Centro Federal de Educação Tecnológica Celso Suckow da Fonseca – (CEFET/RJ)

flavio.carvalho@eic.cefet-rj.br, vanessa.soares@aluno.cefet-rj.br,
ebezerra@cefet-rj.br, gustavo.guedes@cefet-rj.br

Abstract. *This paper evaluates automated feedback generation for ENEM Competency 5 by augmenting zero-shot prompts with an Evaluator Term Set (CTA), a term set proposed in this work and extracted from human evaluator comments on Competency 5. We compare CTA-augmented feedback against a zero-shot baseline on 46 essays with human reference feedback using BERTScore F1 and the Wilcoxon signed-rank test. CTA augmentation increases mean BERTScore F1 for both models, with statistically significant improvement in semantic similarity to human feedback for both evaluated models at* $\alpha = 0.05$.

1. Introduction

The Brazilian National High School Exam (ENEM) is the primary gateway to higher education in Brazil, taken by millions of students annually [INEP, 2024]. Access to consistent feedback is unevenly distributed, with higher-income students benefiting from supplementary courses that provide iterative feedback cycles [Senkevics and Carvalho, 2023; Medeiros, 2016]. Automated feedback can broaden low-income students' access to formative writing support, aligning with Brazil's National Artificial Intelligence Plan (PBIA), which prioritizes AI for social equity [MCTI, 2024].

Automated essay scoring systems assign grades to ENEM essays [Silva and Araujo, 2024]. However, numerical scores alone do not support writing development [Misgna et al., 2024]. Students need feedback that addresses specific competency criteria, yet automated feedback generation remains underexplored [Anchiêta et al., 2025].

Large Language Models (LLMs) enable competency-specific feedback generation via zero-shot prompting [Anchiêta et al., 2025], but are prone to logical inconsistencies [Hooshyar et al., 2025] and hallucinations for underrepresented content [Bezerra, 2025]. Adding domain-specific context to prompts can mitigate some of these limitations [Schreiter, 2025; Bezerra, 2025]. Evaluator comments offer rubric-aligned terms that generic prompts lack.

In this work, we focus on ENEM Competency 5 (intervention proposal), selected for its complexity and underexploration in the automated feedback literature [Anchiêta et al., 2025], and propose an Evaluator Term Set (CTA, *Conjunto de Termos de Avaliadores*), recurring n-grams extracted from human evaluator comments on Competency 5 in the Automated Essay Score (AES) ENEM dataset [Silveira et al., 2024]. We compare CTA-augmented feedback to a zero-shot baseline [Anchiêta et al., 2025] using BERTScore F1 against human reference feedback and assess statistical significance with the Wilcoxon signed-rank test for two locally-run, open-weight LLMs: `mistral_pt`, a Portuguese-adapted model, and `llama32`, a general-purpose multilingual model.

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 78–85
October 19-22, 2026 ©2026 Association for Computational Linguistics

Our contributions include (i) compiling a CTA from human evaluator comments, (ii) proposing a prompt augmentation strategy based on CTA insertion, and (iii) providing empirical evidence that CTA augmentation increases semantic similarity to human feedback with statistical significance for both evaluated models. The paper reviews related work (Section 2), describes our method (Section 3), presents results (Section 4), and discusses conclusions (Section 5).

2. Related Work

We conducted a literature survey in early 2026 using Google Scholar with the terms "LLM", "feedback", and "ENEM". While the search returned over 100 results on automated essay scoring for ENEM, research on feedback generation is sparse. Focusing specifically on feedback generation, we identified two studies that produce textual feedback for ENEM essays using LLM-based systems [Anchiêta et al., 2025; Silva and Araujo, 2024], with one focusing on feedback generation and the other on scoring with feedback as additional output.

Anchiêta et al. [2025] generate feedback for ENEM essays with zero-shot prompting. They evaluate four LLMs on 190 essays from the AES-ENEM dataset [Silveira et al., 2024]. Each prompt has three sections (context, task, output format) and requests one sentence of feedback per competency. They compare generated feedback to human feedback with BERTScore [Zhang et al., 2019], finding similar scores among models (F1 $\approx$ 0.67-0.69). A linguist with ENEM expertise then ranks feedback from 50 essays, identifying Qwen 3-8B as most informative and constructive.

Silva and Araujo [2024] propose an LLM-based solution structured as a prompt chain following the official ENEM rubric [INEP, 2025]. The study focuses on automated scoring and reports performance results for competency-level and final grades using agreement and correlation measures with human evaluators. The system produces competency-specific feedback as supplementary output, described as detailed and criterion-oriented, but does not evaluate feedback quality similarity to human comments.

This work differs from prior approaches in scope and evaluation method. We augment a zero-shot prompt with a CTA extracted from human evaluator comments in a corpus of ENEM essay practice texts, focusing specifically on Competency 5. We compare CTA-augmented feedback against a zero-shot baseline using BERTScore and apply the Wilcoxon signed-rank test to assess whether domain knowledge augmentation increases semantic similarity to human feedback.

3. Methodology

We compare feedback generated by our CTA-augmented approach against the zero-shot baseline proposed by Anchiêta et al. [2025]. We use the `sourceAOnly` subset of the AES-ENEM dataset [Silveira et al., 2024][1], the only subset containing per-competency evaluator comments, drawn from the Educação UOL platform ("Source A") [Silveira et al., 2024] and following a scoring scale consistent with the official ENEM rubric. This subset is divided into train, validation, and test splits.

Of the 258 essays in `train-sourceAOnly`, 167 contain valid Competency 5 feedback, which are used to construct the CTA. The `validation-sourceAOnly`

[1] Available at `https://huggingface.co/datasets/kamel-usp/aes_enem_dataset`.

subset (66 essays, 46 with valid C5 feedback) provides the essay texts for baseline and CTA-augmented feedback generation, and the corresponding human C5 comments serve as reference for BERTScore evaluation. The `test-sourceAOnly` subset is not used in this work. Since all methodological decisions, including CTA construction and parameter selection, were made using the validation subset, the reported results should be interpreted as validation-set evidence, leaving the test set for future confirmation under a fixed experimental design. Figure 1 illustrates our experimental workflow, showing baseline and CTA-augmented paths, CTA construction from evaluator feedback, and evaluation strategy using BERTScore and Wilcoxon testing.

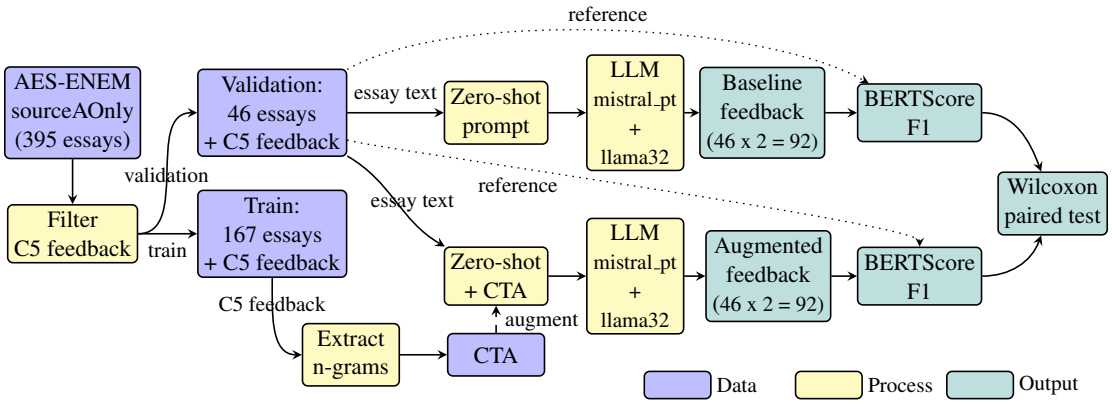


Figure 1. Experimental methodology: CTA construction from training feedback (167 essays), baseline and CTA-augmented generation on the validation subset (46 essays), and BERTScore F1 evaluation with Wilcoxon testing.

Preprocessing begins with the `train-sourceAOnly` subset (258 essays) with fields for prompt, essay text, supporting text, competency scores, and evaluator comments. We extract Competency 5 feedback and retain essays with a valid non-empty C5 comment, yielding 167 essays. The same preprocessing procedure is applied to `validation-sourceAOnly`, yielding 46 essays used for feedback generation and evaluation. Each essay in both subsets is assigned a stable identifier (`essay_hash`) computed as the SHA-256 hash of the normalized essay text, ensuring consistent pairing across experimental conditions.

We adapt the zero-shot prompting strategy described by Anchiêta et al. [2025] to generate baseline feedback. Following the same three-stage structure of context, task, and output format, we translate the prompt to Portuguese to ensure models generate feedback in the essay's language and focus it on Competency 5 only. Figure 2 presents the complete prompt used for baseline generation.

We apply the prompt individually to each essay using two LLMs running locally via Ollama v0.5.7 [Marcondes et al., 2025]. The models are `cnmoro/mistral_7b_portuguese:q2_K` (7B parameters, Q2_K quantization, digest `321345d18e55`; referred to as `mistral_pt`), which was selected for its Portuguese language adaptation, and `llama3.2:latest` (3.21B parameters, Q4_K_M quantization, digest `a80c4f17acd5`, referred to as `llama32`), selected for its general multilingual capability and availability. Both models are configured following Anchiêta et al. [2025], with `top_p`=0.95, `top_k`=50, and temperatures 0.7 for `mistral_pt` and 0.8 for `llama32`. Experiments were executed on an Intel Core i5-10400 CPU with an

NVIDIA GeForce GTX 1650 GPU (4 GB VRAM) running Ollama v0.5.7.

The baseline generates 92 feedback outputs (46 essays $\times$ 2 models). Using the same essays, models, and parameters, the CTA-augmented condition generates another 92 outputs, yielding paired baseline/CTA feedback for each essay. We compare each condition to the human reference feedback with BERTScore F1 and test paired differences with the Wilcoxon signed-rank test.

(a) Zero-shot baseline prompt

Atue como um revisor especializado em redações dissertativo-argumentativas do ENEM.
A redação a seguir foi escrita por um estudante do ensino médio, e você deverá revisá-la em apenas 1 sentença (somente para a Competência 5): {essay_text}
Use o texto de apoio a seguir para verificar se a redação está alinhada ao tema: {supporting_text}
Você deve fornecer feedback apenas para a Competência 5 do ENEM. Na única sentença, analise a proposta de intervenção apresentada na redação. Mantenha um tom construtivo, claro e detalhado.
IMPORTANTE: Siga o formato de saída EXATAMENTE como especificado abaixo. Não adicione qualquer texto adicional, explicações ou formatação fora do formato especificado. Use APENAS o marcador "## Competency 5##" para apresentar o feedback.
Competency 5##
Sua sentença

(b) CTA-augmented prompt (addition highlighted)

Atue como um revisor especializado em redações dissertativo-argumentativas do ENEM.
A redação a seguir foi escrita por um estudante do ensino médio, e você deverá revisá-la em apenas 1 sentença (somente para a Competência 5): {essay_text}
Use o texto de apoio a seguir para verificar se a redação está alinhada ao tema: {supporting_text}

BASE DE CONHECIMENTO — use os termos abaixo como referência avaliativa para enriquecer sua análise. NÃO os repita mecanicamente.
• term_A • term_B … • term_T

Você deve fornecer feedback apenas para a Competência 5 do ENEM. Na única sentença, analise a proposta de intervenção apresentada na redação. Mantenha um tom construtivo, claro e detalhado.
IMPORTANTE: Siga o formato de saída EXATAMENTE como especificado abaixo.
Não adicione qualquer texto adicional, explicações ou formatação fora do formato especificado.
Use APENAS o marcador "## Competency 5##" para apresentar o feedback.
Competency 5##
Sua sentença

Figure 2. Prompt structures used for feedback generation. (a) Zero-shot baseline. (b) CTA-augmented: a knowledge base block is inserted between the essay content and the output instructions; all other content is identical to (a).

The CTA is a knowledge injection strategy, distinct from few-shot prompting. Providing full essay and feedback pairs as in-context examples would require selecting representative instances and risks introducing topic-specific bias, and it also increases prompt length substantially for models with limited context windows. The CTA instead provides a compact, topic-agnostic vocabulary extracted from the evaluator community, compatible with the zero-shot design of the baseline [Anchiêta et al., 2025].

We compile a *Conjunto de Termos de Avaliadores* (CTA, Evaluator Term Set) to augment prompts with domain knowledge specific to Competency 5 (intervention proposal), which is evaluated on five aspects: concrete action, executing agent, means of implementation, detailing, and respect for human rights [INEP, 2025]. The CTA com-

prises recurring n-grams extracted from human evaluator comments on Competency 5 in the AES-ENEM dataset.

We extract n-grams of length 2 to 4 tokens from the 167 human evaluator comments on Competency 5 using the Natural Language Toolkit (NLTK) [Bird et al., 2009], after removing standard Portuguese stopwords. Additionally, we removed 13 high-frequency terms identified as non-discriminative for Competency 5 assessment through empirical frequency analysis of the 167 training comments. These include terms of structural reference (e.g., *autor*, *parágrafo*), topic-specific terms tied to particular essay themes rather than C5 criteria (e.g., *hino*, *armas*), and evaluator format artifacts (e.g., *competência*, *vermelho*). We retain only n-grams present in at least three distinct comments and select the top 20 by document frequency, yielding a single term set of evaluator expressions applied uniformly to all essays. The term extraction procedure results in a ranked list serving as the CTA, and Table 1 shows the 10 highest-ranked terms.

Table 1. 10 highest-ranked terms of the CTA, ranked by document frequency (DF).

Rank	Term	DF	Rank	Term	DF
1	sugestão intervenção	18	6	além disso	5
2	conclusão coerente	16	7	proposta solução	4
3	resolver problema	10	8	sugestões intervenção	4
4	proposta intervenção	6	9	solução problema	4
5	quer dizer	6	10	conclusão confusa	4

The terms reflect expressions tied to the intervention construct (e.g., *sugestão intervenção*, *proposta intervenção*, *resolver problema*) and evaluative quality markers used to signal proposal weaknesses or strengths (e.g., *intervenção genérica*, *conclusão confusa*, *conclusão coerente*). Those two types of evaluator vocabulary provide the LLM with register characteristic of human comments, without prescribing essay content.

We augment the zero-shot prompt with CTA to provide evaluator-grounded context for Competency 5 feedback generation. For each essay, we insert a list of the 20 most frequent terms into the prompt, with one term per line, as shown in Figure 2. The CTA block is inserted between the essay content and the output format instructions, without modifying any other part of the prompt. All 46 essays receive the same CTA block.

We maintain experimental consistency by using the same models (`mistral_pt` and `llama32`) and parameters employed in baseline generation. The CTA block is the only difference between conditions, and all other prompt content and generation parameters are identical. This procedure generates 92 CTA-augmented feedback instances (46 essays $\times$ 2 models) for paired comparison with the baseline.

We evaluate semantic similarity between generated and human reference feedback for Competency 5 using BERTScore F1 [Zhang et al., 2019], measured with `bert-base-multilingual-cased` via the `bert-score` library (`v0.3.13`), setting `lang="pt"` without IDF weighting or baseline rescaling. We conduct pairwise comparisons matching each generated feedback instance with the corresponding human C5 feedback by `essay_hash`, retaining only pairs where both texts exceed 10 characters. BERTScore F1 values are organized by model and condition (baseline $\times$ CTA-augmented) and used as input to the Wilcoxon signed-rank test at $\alpha = 0.05$. All

code, generated feedbacks, CTA term list, and evaluation results are publicly available at `https://github.com/LaCAfe/enem-feedback-cta`, while the AES-ENEM dataset is available separately from its original source [Silveira et al., 2024].

4. Results

Table 2 presents mean BERTScore F1 values for Competency 5, comparing the zero-shot baseline and the CTA-augmented approach for each model. Mean BERTScore F1 values range from 0.673 to 0.686, with standard deviations between 0.020 and 0.028 across models and conditions. The analysis considers only paired comparisons between human and automatic comments for which both conditions produced valid outputs.

Table 2. BERTScore F1 for Competency 5 feedback, comparing the zero-shot baseline and the CTA-augmented approach. Values are mean ± standard deviation over valid paired instances (N = 46) per model.

Model	N	Zero-shot (mean±std)	CTA (mean±std)
`mistral_pt`	46	0.675±0.020	**0.686**±0.028
`llama32`	46	0.673±0.025	**0.683**±0.027

We apply a one-sided paired Wilcoxon signed-rank test (CTA $>$ baseline), justified by the directional hypothesis that CTA augmentation increases semantic similarity to human feedback. Both models show significant improvement at $\alpha = 0.05$: $p = 0.0050$ for `mistral_pt` and $p = 0.0263$ for `llama32` over $N = 46$ paired instances. Bonferroni correction is not applied, as the two models constitute independent replications of the same hypothesis [Perneger, 1998].

To assess whether the observed gains reflect lexical overlap rather than generation quality, we computed BERTScore F1 for synthetic strings formed by concatenating the top-N CTA terms ($N \in \{5, 10, 15, 20\}$) against the same 46 human references. All synthetic conditions scored below the zero-shot baseline (maximum 0.607 ± 0.029 at $N = 5$, versus baseline 0.673 and 0.675), indicating that lexical overlap alone does not account for the CTA-augmented gains.

5. Conclusion

We investigate the use of a *Conjunto de Termos de Avaliadores* (CTA, Evaluator Term Set) as additional knowledge for automated feedback generation on ENEM essays, focusing on Competency 5. We compare a zero-shot baseline with a CTA-augmented approach by measuring semantic similarity between automatic and human feedback using BERTScore and statistical testing. Results indicate that CTA augmentation increases semantic similarity relative to the zero-shot baseline, with both evaluated language models showing statistically significant improvement at $\alpha = 0.05$.

This work suggests that incorporating evaluator-derived terminology from human feedback comments can increase BERTScore semantic similarity between LLM-generated and human feedback, providing a replicable framework for augmenting zero-shot prompts with evaluator content knowledge. Future work should extend evaluation to the remaining ENEM competencies and complement automatic metrics with qualitative analysis by human evaluators to assess feedback qualities not captured by semantic similarity.

Limitations

This study has methodological limitations that should be considered when interpreting results. The quantitative approach using BERTScore as the primary comparison criterion does not capture qualitative aspects of generated feedback such as clarity, pedagogical appropriateness, tone, or actionability. While BERTScore measures semantic similarity to human feedback, it does not assess whether feedback effectively supports student learning or addresses specific writing deficiencies.

Both experimental conditions use a single generation per essay per model. Because LLM outputs are stochastic at temperatures above zero, the observed BERTScore differences reflect one sample from the generation distribution. Multiple independent runs per condition, or fixing the Ollama `seed` parameter alongside the model digest, would be required to confirm the stability of the reported gains; this remains an avenue for future validation.

Regarding scope limitations, our analysis focuses exclusively on Competency 5 (intervention proposal) of ENEM essays. Results may not generalize to other competencies, which assess different writing aspects such as grammatical correctness, argumentation structure, or textual cohesion. The CTA augmentation strategy targets terminology specific to intervention proposals and would require adaptation for other competencies.

The selection of additional stopwords for CTA refinement involved researcher judgment based on empirical frequency analysis. The complete list and the extraction code are publicly available, ensuring full reproducibility of the reported results. However, applying this approach to a different dataset would require a new curation step, as the additional stopwords are specific to the distribution of terms in the training comments used here.

Ethics Statement

The data used in this study come from publicly available sources. No primary data collection involving human participants was conducted. This study analyzes student essays and evaluator comments from two existing public datasets compiled by third parties for research purposes and made available in open repositories and educational platforms.

The research focuses on computational analysis of automated feedback generation for educational assessment. No operational system affecting students' academic outcomes was deployed. Generated feedback was analyzed solely for research comparison against existing human evaluator comments. The study acknowledges that automated methods may reflect biases present in source materials, and results are reported with methodological and scope limitations.

Disclosure of Generative AI Usage

The authors declare the use of generative AI tools during the preparation of this manuscript. Specifically, we used Claude Sonnet 4.5 (Anthropic) to correct the grammar, without changing the tone, style or adding any new information, as well as to assist in writing code followed by revisions and necessary modifications. All final editorial decisions, including verification of consistency, originality, and technical correctness of the content, remained under the sole responsibility of the authors.

References

Anchiêta, R. T., Luz, A. I., Lopes, S. L., and Moura, R. S. (2025). A zero-shot prompting approach for automated feedback generation on ENEM essays. In *Brazilian Symposium on Multimedia and the Web (WebMedia)*, pages 511–515. SBC.

Bezerra, E. (2025). Introduction to LLM-Based agents. In Duarte, D., Timbó, F., Schreiner, G., and Chaves, I., editors, *Tópicos em Gerenciamento de Dados e Informações: Minicursos do SBBD 2025*, chapter 2. Sociedade Brasileira de Computação, Porto Alegre.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Hooshyar, D., Yang, Y., Šír, G., Kärkkäinen, T., Hämäläinen, R., Cukurova, M., and Azevedo, R. (2025). Problems with large language models for learner modelling: Why LLMs alone fall short for responsible tutoring in K–12 education. *arXiv preprint arXiv:2512.23036*.

INEP, I. (2024). ENEM 2024 tem 4,3 milhões de inscritos confirmados. Acesso em: 2 nov. 2024.

INEP, I. (2025). A redação no ENEM 2025: cartilha do(a) participante. Publicado em: 03 out. 2025 08:46. Acesso em: 30 mar. 2026.

Marcondes, F. S., Gala, A., Magalhães, R., Perez de Britto, F., Durães, D., and Novais, P. (2025). Using ollama. In *Natural Language Analytics with Generative Large-Language Models: A Practical Approach with Ollama and Open-Source LLMs*, pages 23–35. Springer.

MCTI, M. (2024). Plano brasileiro de inteligência artificial (PBIA) 2024-2028. Gov.br. Acesso em: 26 set. 2024.

Medeiros, M. (2016). Income inequality in Brazil: new evidence from combined tax and survey data. *World Social Science Report*, page 107.

Misgna, H., On, B.-W., Lee, I., and Choi, G. S. (2024). A survey on deep learning-based automated essay scoring and feedback generation. *Artificial Intelligence Review*, 58(2):36.

Perneger, T. V. (1998). What's wrong with bonferroni adjustments. *Bmj*, 316(7139):1236–1238.

Schreiter, D. (2025). Prompt engineering: How prompt vocabulary affects domain knowledge. *arXiv preprint arXiv:2505.17037*.

Senkevics, A. S. and Carvalho, M. P. D. (2023). Youth and access to higher education: On the non-place of the preparatory course student. *Educação Em Revista*, 39.

Silva, W. A. d. and Araujo, C. C. d. (2024). Automated ENEM essay scoring and feedbacks: A prompt-driven LLM approach. Trabalho de Graduação, Centro de Informática (CIn), Universidade Federal de Pernambuco (UFPE), Recife, Brasil.

Silveira, I. C., Barbosa, A., and Mauá, D. D. (2024). A new benchmark for automatic essay scoring in Portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 1*, pages 228–237.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Empirical Study Towards RBAMR-News: a Rule-Based AMR Parser for News in Portuguese

Maria Julia Bernardo Comarim[1,2], **Ariani Di-Felippo**[1,2,3]

[1]Interinstitutional Center for Computational Linguistics (NILC/USP)

[2]Graduate Program in Linguistics - Federal University of São Carlos (PPGL/UFSCar)

[3]Language and Literature Department - (DL/UFSCar)

mjbcomarim@estudante.ufscar.br, ariani@ufscar.br

Abstract. *This work evaluates AMR graph construction rules, originally developed for literary texts, in Portuguese news data to support building a symbolic AMR parser for this genre. Using the AMRNews-PT gold-standard corpus, we conduct an ablation analysis to assess rule contributions to AMR representation quality. The results reveal a hierarchical organization in which a small set of core structural rules drives most parsing performance, while additional rules provide finer semantic refinements. This rule hierarchy serves as the basis for building the parser and enabling the annotation of a larger news corpus.*

1. Introduction

Semantic representation refers to structured models of natural language meaning. This capability is essential to enhance machine understanding of human language [Jurafsky and Martin, 2026]. Among deep semantic formalisms, Abstract Meaning Representation (AMR) [Banarescu et al., 2013] emerges as a versatile framework, proving effective in diverse applications such as machine translation, summarization, and information extraction [Wein and Opitz, 2024]. The success of AMR in Natural Language Processing (NLP) can be attributed to three main factors: (i) its abstract, graph-based semantic representation that abstracts away from surface syntax; (ii) its standardized annotation scheme with clear guidelines[1] ensuring consistency and enabling gold-standard corpora; and (iii) its grounding in PropBank frame semantics [Palmer et al., 2005], providing a stable lexical-semantic foundation [Sadeddine et al., 2024; Mansouri, 2025].

The most important corpus is the AMR Annotation Release [Banarescu et al., 2013]. It was constructed fully manually by trained annotators, and contains about 60,000 sentences in its latest (3.0) version. This resource has enabled the development of monolingual[2] AMR parsing across multiple modeling paradigms, fostering state-of-the-art (SOTA) transformer-based models [Bevilacqua et al., 2021; Zhou et al., 2021; Bai et al., 2022; Vasylenko et al., 2023] and competitive approaches based on parameter-efficient strategies [Ho, 2025]. However, these approaches remain highly data-dependent, relying heavily on large annotated corpora such as AMR 3.0 for training and performance gains. This limits their applicability to low-resource settings such as Brazilian Portuguese (BP), whose largest available corpus, AMR-LittlePrince [Anchiêta and Pardo, 2018], comprises around 1.5k sentences and is restricted to a specific literary genre (allegorical narrative).

[1]https://github.com/amrisi/amr-guidelines/blob/master/amr.md

[2]This work focuses exclusively on monolingual AMR parsing, enabling direct modeling of language-specific phenomena without reliance on cross-lingual transfer from English resources.

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 86–99
October 19-22, 2026 ©2026 Association for Computational Linguistics

To address this gap, we focus on approaches that provide stronger structural control while reducing dependence on large annotated corpora, aiming to support reliable AMR parsing in under-resourced languages. Although large language models (LLMs) with in-context learning (ICL) strategies have recently emerged as a promising alternative for low-resource AMR parsing, results suggest that they are not yet reliable parsers [Ettinger et al., 2023; Li and Fowlie, 2025]. Consequently, symbolic approaches remain relevant due to their interpretability and explicit structural modeling, particularly in settings such as Portuguese news AMR parsing.

Thus, we aim to develop RBAMR-News, a symbolic AMR parsing method for the Portuguese news genre. To this end, we take RBAMR (which stands for "Rule-Based AMR Parsing") [Anchiêta and Pardo, 2018] as a starting point. It is a pipeline-based architecture composed of syntactic parsing, semantic role labeling (SRL), and rule-based AMR generation modules. This symbolic method was originally built on the AMR-LittlePrince corpus and has not previously been evaluated on the news genre. As the original parser is no longer executable, we reconstruct its pipeline scheme using recently available tools employed independently, without pipeline-level integration. This reconstruction enables a broader four-stage framework for developing RBAMR-News: (i) evaluation of the original RBAMR rules on news text, (ii) coverage expansion through the investigation of genre-specific phenomena and the proposal of new rules from a larger corpus annotated with gold-standard UD and PropBank semantic roles, (iii) iterative diagnostic projection and rule refinement, and (iv) final validation of the resulting rule hierarchy.

In the present paper, we focus on the reconstruction of RBAMR pipeline scheme and on the first stage of the proposed framework, assessing the transferability of its original rule inventory to the news domain through an ablation-based analysis on AMRNews-PT [Sobrevilla Cabezudo and Pardo, 2019], the only news corpus with manual AMR annotation, containing 870 sentences. The remainder of this article is structured as follows. Section 2 briefly introduces AMR fundamentals, while Section 3 provides the main related work. Section 4 presents the reconstruction of the RBAMR parser, detailing the adaptation of each component, especially the original rules, to the current setup. Section 5 reports the experimental design, including the dataset and evaluation metrics. Section 6 summarizes the main findings and outlines directions for future work.

2. Abstract Meaning Representation

AMR is a deep semantic formalism characterized by graph-based representation, abstractness, and uniformity. It represents sentence meaning as rooted, directed, acyclic graphs. (Figure 1(c))[3]. Nodes of an AMR graph are concepts, and edges are relationships between those concepts. AMR concepts are either English lexical items (e.g., "boy"), PropBank frames [Palmer et al., 2005] (e.g., *desire-01*), or specialized keywords, which include special named entity types, quantities, and logical operators. Leaf nodes are labeled with semantic concepts, such that *(b / boy)* denotes an instance of the concept *boy*, where the variable *b* serves as the identifier of the node in the graph. For example, in *(d / desire-01 :ARG0 (b / boy))*, the variable *d* introduces an event node instantiating the PropBank frame *desire-01*, which is linked via the *:ARG0* edge to the variable *b*, corresponding to the entity *boy*. This structure encodes that *b* fills the *:ARG0* role (wanter) of the *desire-*

[3]The PENMAN notation (Figure 1(b)) is used as a linearization format for human reading and writing.

01 event, establishing a semantic relation between event and participant. The roleset is derived from the PropBank frame repository [Palmer et al., 2005]. In the case, the frame *desire-01* has a roleset with two pre-defined arguments (:arg0=wanter and :arg1=wanted), which are filled by *boy* and *believe-01* in the sentences shown in Figure 1.

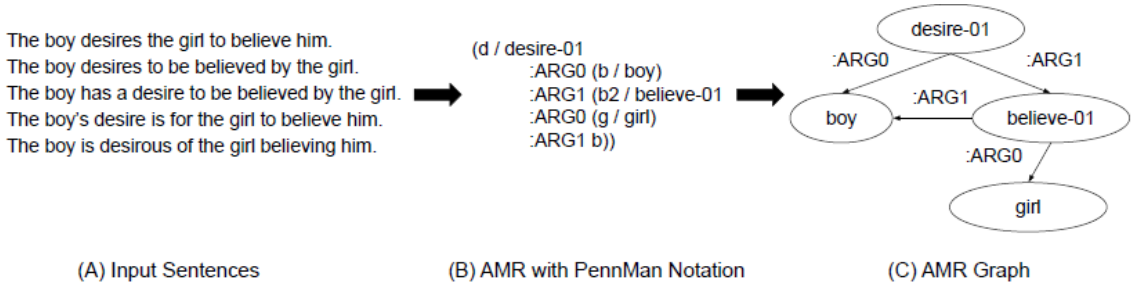


Figure 1. Example of AMR formalism [Mansouri 2025].

Edges show the relations between nodes in the graph, supporting compositional meaning construction. When a single participant fulfills multiple semantic roles, AMR encodes this via reentrancy (nodes with multiple parents) or variable re-use in PENMAN notation. The AMR relation inventory comprises approximately one hundred semantic relations, grouped into functional classes: (i) frame arguments (PropBank roles), (ii) general relations, (iii) quantities, (iv) date-entities, and (v) relations for lists. AMR also defines inverse relations (e.g., *:location-of*), which allow reverse role interpretation within the graph structure. Additionally, each relation admits a reified counterpart, enabling the relation itself to be treated as a first-class semantic entity; for instance, *:location* can be reified as *be-located-at-91* [Banarescu et al., 2013].

Another principle is to abstract away from syntactic structure, meaning that sentences with different word orders or syntactic constructions conveying the same meaning will have the same AMR (Figure 1(a)). This abstraction facilitates better generalization for downstream tasks. Note that functional words like “as” or “it” are not annotated, as they are considered not to contribute additional semantic content. AMR also aims to provide a uniform, consistent semantic representation for different linguistic constructions. For example, both active and passive voice sentences have the same AMR as long as they convey the same meaning, regardless of differences in syntactic structure.

AMR annotation follows a set of guidelines originally introduced by Banarescu et al. [2013] and subsequently refined to incorporate a broader range of linguistic phenomena, and increase their application for downstream tasks. Efforts to develop annotation guidelines for AMR have extended to several non-English languages. For Portuguese, these efforts adapt the model to language-specific phenomena such as null subjects, diminutives, possessive pronoun ambiguity, light verbs, and idiomatic expressions [Inácio et al., 2023], as well as to the specificity of stock market tweets [Ceregatto, 2025]. In the following section, we present the reengineering of RBAMR, focusing mainly on the adaptation of its components to evaluate their suitability for the news genre.

3. Related Work

The foundational AMR work introduced in 2013 [Banarescu et al., 2013] established the annotation guidelines and graph-based semantic representation framework, leading to manually annotated corpora (including *The Little Prince*) and the subsequent release of

large-scale English AMR datasets by the Linguistic Data Consortium[4] (LDC). The LDC has released three successive versions of AMR annotations and datasets, with each release incorporating updates and refinements over previous versions. These datasets are AMR 1.0, 2.0 e 3.0, with 13,051, 39,260 and 59,255 sentences, respectively. AMR 3.0 is the most comprehensive, with the largest number of sentences and the most refined and consistent annotations, including natural language sentences from broadcast conversations, newswires, weblogs, web discussion forums, fiction, and web text.

These resources have enabled the development of monolingual parsers across a variety of NLP paradigms, with Transformer-based approaches driving the most significant advances [Mansouri, 2025]. Recent systems rely on full fine-tuning of seq-to-seq models such as BART, formulating AMR parsing as text-to-graph generation via graph linearization, and have substantially improved Smatch performance [Bevilacqua et al., 2021; Zhou et al., 2021; Bai et al., 2022; Vasylenko et al., 2023]. For example, LeakDistill [Vasylenko et al., 2023] achieves Smatch scores of 0.861 on AMR 2.0 and 0.856 on AMR 3.0, positioning it as a current SOTA AMR parser. More recently, parameter-efficient fine-tuning (PEFT) methods such as LoRA have been applied to large language models (such as Llama 3.2) for AMR parsing, reducing adaptation costs while maintaining competitive performance (0.80) [Ho, 2025]. In parallel, researchers have begun investigating the ability of LLM to handle AMR parsing through ICL [Ettinger et al., 2023; Li and Fowlie, 2025]. The results indicate that, despite generating well-formed and partially correct AMR structures, LLMs are still not reliable out-of-the-box AMR parsers, with maximum Smatch scores reaching only around 0.60 with GPT-4o in a 5-shot setting.

Given this, top-performance approaches rely on large annotated corpora and high computational cost, and they inherit limitations from linearization-based strategies, particularly in representing reentrancies and long-distance dependencies. These constraints are even more critical in low-resource languages, where the scarcity of annotated data further limits the applicability of large-scale neural models. AMR resources[5] for Portuguese remain limited in size and domain coverage (Table 1), with the largest corpus, AMR-LittlePrince [Anchiêta and Pardo, 2022] (1,527 sentences), restricted to literary text. All the corpora were manually or semi-automatically (with human revision) annotated based on Verbo-Brasil[6] [Duran and Aluísio, 2012; Sanches Duran and Aluísio, 2015] - a repository of verbs and frames inspired by the English PropBank.

Table 1. Corpora with AMR annotation in Portuguese.

Corpus	Genre	Size	Reference
AMR-LittlePrince	Literature	1,527	Anchieta and Pardo (2018a)
DANTESTocks-AMR	Tweet	1,128	Ceregatto and Di-Felippo (2025)
AMRNews-PT	News	870	Sobrevilla Cabezudo and Pardo (2019)
OpiSums-PT-AMR	Online review	404	Inácio et al. (2023)
AMRScien-Br-Corpus	Popular science	200	Seno et al. (2022)

This scarcity of annotated resources is also reflected in the limited availability of AMR parsers. For Portuguese, three main parsers have been evaluated on the manually anno-

[4]https://catalog.ldc.upenn.edu/

[5]https://github.com/nilc-nlp/AMR-BP

[6]http://143.107.183.175:21380/verbobrasil/

tated AMR-LittlePrince corpus. RBAMR is a rule-based parser that leverages syntactic and semantic information (SRL) to construct AMR graphs. Its rules were derived from AMR-LittlePrince, and when tested on the same dataset, it achieved the best results, with an average Smatch F-score of 0.535. Later, with the addition of new rules and a pruning method, its average Smatch increased to 0.57. CAMR [Wang et al., 2015] and AMREager [Damonte et al., 2017] are transition-based parsers trained and tested on the same corpus, adapted to Portuguese with pre-processing tools such as Stanza, LX-Parser, and spaCy, as well as translated lexical resources and pretrained Portuguese embeddings [Hartmann et al., 2017]. CAMR converts dependency trees into AMR graphs through aligned transitions and reached an average Smatch F-score of 0.455, while AMREager processes sentences incrementally from left to right and obtained 0.445.

Thus, advancing AMR parsing for Portuguese requires not only improvements in parsing methods, but also a substantial expansion of annotated corpora. In its low-resource setting, rule-based approaches remain relevant for the semi-automatic construction of gold-standard AMR corpora, as they provide controlled and transparent initial representations that facilitate manual revision and improve annotation consistency. Hence, we are currently developing the RBAMR-News parser to the news genre, which is underrepresented in the AMR resources (Table 1). Here, we investigate the portability of the pioneering RBAMR parsing rules to news texts, aiming to assess its adequacy for supporting semi-automatic AMR corpus construction in this genre. In the next section we introduce AMR and explain its fundamental concepts and structures.

4. RBAMR Reengineering

The RBAMR parser was implemented as a pipeline architecture composed of three modules: (i) syntactic parsing (Palavras [Bick, 2000]), (ii) SRL (Brazilis [Hartmann et al., 2016] based on [Alva-Manchego and Rosa, 2012]), and (iii) rule-based generation. Since the parser relies on outdated libraries and legacy dependencies, the system had to be reengineered to support this experiment. To preserve the original architecture as closely as possible, we replaced each module with contemporary tools providing analogous functionality, although without fully restoring the original end-to-end integration (Figure 2).

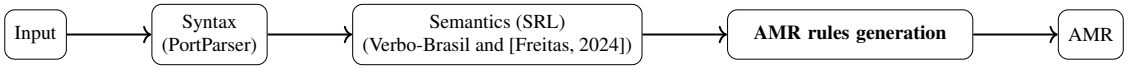


Figure 2. Pipeline-inspired architecture for rule evaluation.

For the syntactic module, we used PortParser v.2 [Lopes and Pardo, 2024]. Built upon the LatinPipe architecture [Straka et al., 2024], the parser achieves the best performance for Portuguese under the Universal Dependencies (UD) framework [de Marneffe et al., 2021], reaching 97.17% UAS (Unlabeled Attachment Score) and 96.19% LAS (Labeled Attachment Score). It was trained on the Porttinari-base v2 corpus[7] using BERTimbau-*large* [Souza et al., 2020] (*bert-base-portuguese-cased*) embeddings. The output of the parser consists of sentences in CoNLL-U format, containing token-level morphological and syntactic dependency annotations compliant with the UD formalism.

[7]https://sites.google.com/icmc.usp.br/poetisa/porttinari-2-1

For the SRL component, we consider BERTimbau (large) trained on PropBank.Br[8] (first version - v1.1)[9], since it is among the best-performing SRL models for Portuguese (see Oliveira et al. [2021]). However, the model requires predicate positions to be explicitly provided as input and produces argument annotations as flat BIO label sequences. Both characteristics pose practical limitations for the experimental setting and would require substantial architectural reformulation for integration into our approach.

Therefore, a heuristic-based symbolic SRL module was developed to operate directly on UD annotations by first identifying verbal predicates and their syntactic arguments from dependency relations such as *nsubj*, *obj*, and *obl*. Next, the module consults Verbo-Brasil to select the most compatible roleset according to the extracted verbal valency. Semantic roles are assigned by mapping UD dependency relations to semantic arguments, with *nsubj* and *obj* generally mapped to *Arg0* and *Arg1*, respectively, accounting for passive constructions Freitas [2024]. The module also handles implicit subjects in subordinate and coordinated clauses through recursive dependency-tree traversal. When a predicate is not found in Verbo-Brasil, a deterministic fallback strategy generates a generic roleset (*lemma.01*), preserving the continuity of the semantic representation. In the end, the SRL returns the predicates with their respective arguments. The SRL module was evaluated on the Porttinari-Base PropBank (PBP) corpus[10], comprising 13,395 verbal predicates annotated with semantic roles. Our module achieved an F1-score of 0.413. By comparison, the supervised SRL system originally employed by RBAMR obtained an F1-score of 0.673 on PropBank [Alva-Manchego and Rosa, 2012]. Although this provides a useful point of reference, the two scores are not directly comparable because they were obtained on different corpora and under different evaluation protocols.

Particularly relevant to this work are the handcrafted rules derived from the manual annotation of the AMR-LittlePrince corpus, which are responsible for constructing the final AMR graph from the intermediate representations. The adoption of the PortParser required the reformulation of the rules, adapting surface-based patterns into syntactic dependency structures compatible with its output. Both the original rules and their adapted versions used in the experiment are presented in Table 2. We exemplify the adaptation with the direct object rule. In the original parser, direct objects marked with the ACC (accusative) tag by PALAVRAS were mapped to the AMR relation *:ARG1*. In our experiment, we adapted this rule so that the UD dependency relation *obj* is mapped to *:ARG1*.

The rules can be broadly organized into four main categories according to the linguistic level they operate on. Lexical and morphosyntactic rules operate over lexical units, POS tags, and morphological features to support AMR concept generation, auxiliary detection, implicit subject recovery, and functional word filtering. Named-entity rule generates standardized AMR subgraphs for entities through canonical *:name* structures and their corresponding *:opx*[11] children. Argument-structure rules are driven by the output of SRL, mapping Verbo-Brasil semantic roles to core AMR relations (e.g., *:ARG0*, *:ARG1*)

[8]PropBank.Br is the (Brazilian) Portuguese version of PropBank and it follows the same annotation style as the original, having a similar role set [Duran and Aluísio, 2012]

[9]It contains 5,931 manually annotated instances covering 3,348 sentences from the Bosque Treebank [Afonso et al., 2002], which is derived from the 1994 *Folha de São Paulo* newspaper.

[10]https://sites.google.com/icmc.usp.br/poetisa/porttinari-base-propbank

[11]The named entity "Nobel Prize", for example, is represented through a `:name` structure whose ordered tokens are encoded by the relations *:op1* "Nobel" and *:op2* "Prize".

while also resolving reentrancies and shared arguments. Finally, syntactic-semantic rules transform dependency relations into AMR edges, handling phenomena such as modification, negation, quantification, and copular inversions. For adverbial modification, we employ a lexical heuristic to refine the generic UD relation *advmod* produced by PortParser into more semantically specific AMR relations, such as *:time*, *:degree*, and *:manner*.

Table 2. Original rules and adaptations.

Phenomenon	Original Rules	Adaptations
Lexical and Dictionary-Based Rules		
Auxiliary detection	PALAVRAS (AUX) labels	UD AUX + Hartmann's rules
Subject insertion	Morphology (Person/Number)	Person/Number→virtual pronoun
Ordinals	Lexical ordinal rules	ADJ+°/ª→ordinal-entity
Stopword filtering	Functional word list	Stopword list filtering
Named Entity Rule	PROPN+Gazetteers (Gaz.)	PROPN+Gaz.→named AMR node
Argument Structure Rules		
Subject	SUBJ → :ARG0	nsubj → :ARG0
Passive subject	v-pcp + SUBJ → :ARG1	nsubj:pass → :ARG1
Direct object	ACC → :ARG1	obj → :ARG1
Indirect object	DAT → :ARG2	iobj → :ARG2
Open clausal complement	Infinitival (IV) rules	xcomp → :ARG2
Nominal modifier	Nominal dependency rules	nmod → :ARG1
Syntactic-Semantic Rules		
Adjectival modifier	ADJ attachment rules	amod → :mod
Adverbial modifier	ADV rules	advmod → :time, :degree, :manner
Coordination	Coordination (KC) → :op	cc/conj → :op
Quantification	NUM attachment rules	nummod → :quant
Possession	Pronoun/Genitive → :poss	nmod:poss → :poss
Negation	ADV(*não/jamais*)→:polarity -	*não/nunca*→ :polarity -
Obligation modality	Aux (*dever*)→ obligate-01	*ter/dever/haver*→ :obligate
Possibility modality	Aux (*poder*)→possible-01	*poder*→ :possible
Passive voice	Pattern *ser*+v-pcp	aux:pass
Copular constructions	*ser/estar*→be-01	*ser/estar*→be-03
Contrast	Discourse (*mas/porém*)	*mas/porém/todavia*→:contrast
Questions	Punctuation (?)+interrogatives	Interrogative sent. heuristics

5. RBAMR Rules Evaluation on News Texts

The RBAMR rules were evaluated using the AMRNews-PT corpus [Sobrevilla Cabezudo and Pardo, 2019] (see Table 1), whose texts were extracted from different sections of the Folha de São Paulo news agency, and also from PropBank.Br. The corpus contains 870 manually annotated sentences, with length limited to up to 23 tokens[12]. The sentences were annotated based on the original English AMR guidelines with adaptations for Portuguese and using Verbo-Brasil. The annotation was carried out by trained annotators from Computer Science and Linguistics. The process involved four steps: (i) identification of sentence type to handle complex structures (e.g., coordination and subordination); (ii) concept identification, (iii) semantic role assignment, and (iv) AMR graph construction using the AMR Editor tool [Hermjakob, 2013][13]. Using a set of 50 sentences, the overall agreement for AMRNews-PT achieved a Smatch [Cai and Knight, 2013] value of 0.73, which is a good value, considering that the IAA in the original AMR project ranged between 0.70 and 0.80.

[12]Due to the complexity of annotation, the corpus includes only sentences shorter than the average length.

[13]AMR Editor is no longer available and has since been replaced by metAMoRphosED [Heinecke, 2023]

The evaluation is conducted at two complementary levels. At the rule level, we first assess the relevance of the symbolic AMR rules in the target journalistic corpus using support, defined as the mean number of rule applications per sentence. Unlike the classical definition of support in association rule mining, which represents a relative frequency bounded by 1 [Agrawal and Srikant, 1994], this metric measures rule application frequency and may therefore exceed 1 when a rule is applied multiple times within the same sentence.

This is particularly important for evaluating the transferability of rules originally designed for a different genre. We then assess rule performance using precision, recall, and F1-score, following standard definitions in information extraction [Jurafsky and Martin, 2026]. Precision is defined as the proportion of correct rule applications among all triggered instances (TP/(TP+FP)), while recall measures the proportion of correctly identified instances among all relevant cases in the corpus (TP/(TP+FN)). The F1-score corresponds to the harmonic mean of precision and recall, providing a balanced measure between correctness and coverage. This stage enables a fine-grained analysis of both the accuracy and coverage of each rule in capturing relevant linguistic patterns. At the (pipeline-inspired) method level, we evaluate the quality of the resulting AMR structures using Smatch, the standard metric for AMR graph comparison [Cai and Knight, 2013], which computes the F-score over concept and relation overlaps between generated and gold-standard graphs. To further assess the contribution of individual rules to the overall system performance, we introduce an ablation-based analysis, which is described in detail in the following subsection.

We apply an ablation-based evaluation [Meyes et al., 2019] to quantify the impact of each symbolic AMR rule on the overall pipeline-inspired parsing, identifying the rules most relevant for semantic representation and cross-genre generalization. The procedure consists of iteratively removing one rule at a time from the system while keeping all other components unchanged. After each removal, the full AMR parsing "pipeline" is re-executed on the evaluation corpus. Performance is then measured using Smatch at the system level, as well as precision and recall where applicable. To quantify the contribution of each rule, We compute the performance delta for each ablation as:

$$\Delta S = S_{\text{ablated}} - S_{\text{full}}, \quad \Delta P = P_{\text{ablated}} - P_{\text{full}}, \quad \Delta R = R_{\text{ablated}} - R_{\text{full}}$$

The equations represent the performance variation induced by each ablation setting. Specifically, ΔS, ΔP, and ΔR correspond to the differences in Smatch, precision, and recall, respectively, between the ablated version (ablated) and the complete system (full). Positive values indicate that removing a module improves performance, suggesting it is detrimental to the system, while negative values indicate that removing it degrades performance, suggesting it is beneficial.

For the evaluation of the rules, support was calculated over the entire corpus, while the ablation experiments were conducted on a randomly selected subset of 435 sentences. This decision was motivated by the empirical stability observed in preliminary experiments, suggesting that the subset provided a representative approximation of the parser's overall behavior, while also reducing processing time and facilitating iterative inspection of the linguistic heuristics. Additionally, some ablation settings produced structurally inconsistent graphs, making large-scale one-to-one comparisons for precision, recall, and Smatch calculations less reliable.

6. Results and Discussion

Table 3 reports the Smatch scores obtained by comparing the output of our method against the gold-standard AMR annotations of the AMRNews-PT corpus. These results serve as the baseline for our experiments, as they correspond to the full configuration in which all rules are activated, providing a reference point for the subsequent ablation analysis. The overall performance is shown in Table 3a, where the system achieves a Smatch F1-score of 0.51, with precision of 0.43 and recall of 0.64. The fine-grained analysis in Table 3b indicates that concept identification (0.57 F1) is substantially more reliable than relation prediction (0.39 F1), reflecting the current dependency of the model on UD-based mappings for semantic role assignment. Compared to the Smatch F1-score of 0.535 obtained by the original RBAMR on the AMR-LittlePrince corpus, our pipeline-inspired method achieves similar performance. Table 3 summarizes the overall behavior of the full system without isolating the contribution of each symbolic rule. In order to address this aspect, we introduce an ablation study. For comparison, we also computed Smatch scores on a 435-sentence subset of the corpus used in the ablation study, as indicated in Table 4.

The ablation results reveal a stratification of rule importance in the news corpus. Core semantic backbone rules—subject (*nsubj*→:ARG0), direct object (*obj*→:ARG1), and named entities (highlighted in blue in Table 4)—consistently yield the largest drops in Smatch (ΔS) when removed, indicating their central role in preserving predicate–argument structure and ensuring faithful entity grounding in the resulting AMR graphs. In contrast, a set of grammatical phenomena (highlighted in orange), including copular constructions, possession, ordinals, and indirect objects, shows negligible variation across Smatch, precision, and recall (ΔS, ΔP, $\Delta R \approx 0$), in some cases with zero corpus support, suggesting that their contribution to global semantic reconstruction is minimal. An intermediate group (highlighted in purple), comprising adverbial modification, nominal modification, and coordination, exhibits small but consistent effects ($\Delta S \sim 0$ to -0.02), reflecting a secondary role in refining event semantics rather than defining core structure. Finally, auxiliary-related phenomena cannot be evaluated in isolation due to interaction effects with other rules, indicating non-modular dependencies within the system.

Table 3. Pipeline-inspired method evaluation on AMRNews-PT.

(a) Overall performance

Metric	Value
Precision	0.43
Recall	0.64
F1-score	**0.51**

(b) Fine-grained performance

Category	F1-score
Concepts	0.56
Relations	0.39

7. Final Remarks and Future Work

The ablation results provide a hierarchy in which core structural rules are prioritized, intermediate rules support semantic refinement, and low-impact rules are selectively included or pruned according to performance–cost trade-offs. Future work will focus on expanding the rule hierarchy through the investigation of genre-specific phenomena that may require refinements to existing AMR treatments or even the proposal of new annotation guidelines, which may in turn motivate the formulation of additional rules. This process will be conducted based on the Porttinari-Base PropBank (PBP) corpus [Freitas

Table 4. Ablation experiment for each rule.

Phenomenon	Supp.	Smatch	ΔS	Prec.	ΔP	Rec.	ΔR
Baseline (435 sents)	—	**0.52**	—	**0.43**	—	**0.64**	—
Stopword filtering	1.4023	0.52	0.00	0.44	+0.01	0.63	-0.01
Named entities	0.5080	0.47	-0.05	0.38	-0.05	0.61	-0.03
Subject (nsubj → :ARG0)	0.4494	0.46	-0.06	0.36	-0.07	0.65	+0.01
Auxiliary detection	0.3391	—	—	—	—	—	—
Direct object (obj → :ARG1)	0.3218	0.47	-0.05	0.37	-0.06	0.65	+0.01
Nominal modification (nmod → :ARG1)	0.3000	0.50	-0.02	0.40	-0.03	0.67	+0.03
Subject insertion	0.2506	0.50	-0.02	0.40	-0.03	0.65	+0.01
Adjectival mod. (amod → :mod)	0.2241	0.51	-0.01	0.41	-0.02	0.65	+0.01
Manner (advmod → :manner)	0.1368	0.52	0.00	0.43	0.00	0.65	+0.01
Time (advmod → :time)	0.0414	0.51	-0.01	0.43	0.00	0.64	0.00
Degree (advmod → :degree)	0.0345	0.52	0.00	0.43	0.00	0.64	0.00
Coordination (cc/conj → :op)	0.1621	0.52	0.00	0.43	0.00	0.66	+0.02
Questions	0.0506	0.52	0.00	0.43	0.00	0.64	0.00
Negation (não/nunca)	0.1379	0.51	-0.01	0.42	-0.01	0.64	0.00
Open clausal comp. (xcomp → :ARG2)	0.1023	0.52	0.00	0.44	+0.01	0.63	-0.01
Quantification (nummod → :quant)	0.0805	0.52	0.00	0.43	0.00	0.65	+0.01
Passive subject (nsubj:pass → :ARG1)	0.0368	0.51	-0.01	0.42	-0.01	0.64	0.00
Passive voice (aux:pass)	0.0000	0.52	0.00	0.43	0.00	0.64	0.00
Contrast (:contrast)	0.0287	0.52	0.00	0.44	+0.01	0.63	-0.01
Obligation (ter/dever)	0.0195	0.52	0.00	0.43	0.00	0.64	0.00
Possibility (poder)	0.0161	0.52	0.00	0.43	0.00	0.64	0.00
Copular constructions (ser/estar)	0.0011	0.52	0.00	0.43	0.00	0.64	0.00
Ordinals (ordinal-entity)	0.0011	0.52	0.00	0.43	0.00	0.64	0.00
Indirect object (iobj → :ARG2)	0.0000	0.52	0.00	0.43	0.00	0.64	0.00
Possession (nmod:poss → :poss)	0.0000	0.52	0.00	0.43	0.00	0.64	0.00

and Pardo, 2024, 2025], which comprises 8,418 sentences annotated with gold-standard UD and PropBank structures. The resulting rule hierarchy, expanded with genre-specific rules, will then be applied to this corpus to automatically generate draft AMR graphs for manual inspection and correction, simultaneously supporting iterative parsing refinement and the semi-automatic construction of additional AMR-annotated data. Finally, the expanded rule hierarchy will be reevaluated on AMRNews-PT through controlled ablation experiments to assess the effectiveness of the proposed adaptation strategy, ultimately resulting in a fully developed RBAMR-News parser. In the broader context of Portuguese, where annotated AMR resources remain scarce and domain coverage is still limited, this strategy may contribute to the incremental construction of larger AMR corpora (PBP), while providing empirically grounded insights into the adaptation of symbolic parsing frameworks across textual genres[14].

Acknowledgements. This work was carried out at the Center for Artificial Intelligence of the University of São Paulo (C4AI - http://c4ai.inova.usp.br/), with support by the São Paulo Research Foundation (FAPESP #2019/07665-4) and by the IBM Corporation. The project was also supported by the Ministry of Science, Technology and Innovation, with resources of Law N.8.248, of Oct 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published as Residence in TIC 13, DOU 01245.010222/2022-44.

Limitations

A current limitation of our experimental setup is the reliance on a rule-based SRL component and the Verbo-Brasil lexicon within the RBAMR pipeline, which constrains the

[14]Experiments, resources and other information related to this paper are available at: https://github.com/mjbcomarim/RBAMR-News

quality of the intermediate semantic representation used for AMR construction. Since the AMR rules operate over SRL outputs, variations in SRL quality can directly affect argument structure identification and, consequently, the downstream semantic graph generation. In particular, this dependency may influence the stability and relative importance of rules when applied to more accurate SRL models. To address this limitation, future work will replace the current SRL module with a higher-quality neural SRL system (e.g., BERT-based models) and re-evaluate the ablation-based rule analysis under this updated intermediate representation. This will allow us to assess the robustness of the proposed rule hierarchy to improvements in semantic role prediction and to quantify the extent to which rule effectiveness is dependent on SRL quality versus intrinsic structural relevance in the news domain.

Ethics Statement

This work relies on previously available NLP tools, including syntactic parsing and SRL rule-based strategies, and linguitic resources, particularly news-domain corpora with gold-standard syntactic and semantic annotations. No personal or sensitive data were intentionally collected or processed. As this work involves NLP tools and rule-based AMR parsing, potential biases and error propagation effects originating from upstream syntactic and SRL may propagate to the generated AMR structures. Nevertheless, we expect the proposed methodology to contribute positively to the expansion of semantic resources and symbolic parsing strategies for Portuguese, a language still underrepresented in AMR research.

Disclosure Regarding the Use of AI Tools

Using GenAI only to correct the grammar, without changing the tone, style or adding any new information.

References

Afonso, S., Bick, E., Haber, R., and Santos, D. (2002). Floresta sintá(c)tica: A treebank for portuguese. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands, Spain. European Language Resources Association (ELRA).

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, page 487–499, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Alva-Manchego, F. E. and Rosa, J. L. G. (2012). Semantic role labeling for brazilian portuguese: A benchmark. In *Advances in Artificial Intelligence – IBERAMIA 2012*, volume 7637 of *LNCS*, pages 481–490, Berlin, Heidelberg. Springer.

Anchiêta, R. T. and Pardo, T. A. S. (2018). A rule-based amr parser for portuguese. In Simari, G. R., Fermé, E., Gutiérrez Segura, F., and Rodríguez Melquiades, J. A., editors, *Advances in Artificial Intelligence - IBERAMIA 2018*, pages 341–353.

Anchiêta, R. T. and Pardo, T. A. S. (2022). Abstract meaning representation parsing for the brazilian portuguese language. In Pinheiro, V., Gamallo, P., Amaro, R., Scarton, C., Batista, F., Silva, D., Magro, C., and Pinto, H., editors, *Computational Processing of the Portuguese Language*, pages 429–434, Cham. Springer International Publishing.

Bai, X., Chen, Y., and Zhang, Y. (2022). Graph pre-training for AMR parsing and generation. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6001–6015, Dublin, Ireland. Association for Computational Linguistics.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N. (2013). Abstract Meaning Representation for sembanking. In Pareja-Lora, A., Liakata, M., and Dipper, S., editors, *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Bevilacqua, M., Blloshmi, R., and Navigli, R. (2021). One spring to rule them both: Symmetric amr semantic parsing and generation without a complex pipeline. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12564–12573.

Bick, E. (2000). *The Parsing System Palavras: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press, Aarhus.

Cai, S. and Knight, K. (2013). Smatch: an evaluation metric for semantic feature structures. In Schuetze, H., Fung, P., and Poesio, M., editors, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752, Sofia, Bulgaria. Association for Computational Linguistics.

Ceregatto, G. (2025). Representação formal de significado: o caso dos tweets do mercado financeiro. Dissertação (mestrado em linguística), Universidade Federal de São Carlos, São Carlos.

Damonte, M., Cohen, S. B., and Satta, G. (2017). An incremental parser for abstract meaning representation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 536–546.

de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Duran, M. S. and Aluísio, S. M. (2012). Propbank-br: a Brazilian treebank annotated with semantic role labels. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1862–1867, Istanbul, Turkey. ELRA.

Ettinger, A., Hwang, J., Pyatkin, V., Bhagavatula, C., and Choi, Y. (2023). "you are an expert linguistic annotator": Limits of LLMs as analyzers of Abstract Meaning Representation. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8250–8263, Singapore. Association for Computational Linguistics.

Freitas, C. (2024). Anotação de papéis semânticos no corpus porttinari-base: procedimentos, resultados e análises. Technical Report 450, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos-SP.

Freitas, C. and Pardo, T. A. S. (2024). Propbank e anotação de papéis semânticos para a língua portuguesa: O que há de novo? In *Proceedings of the 15th Symposium in Information and Human Language Technology (STIL)*, pages 118–128, Belém-PA, Brazil.

Freitas, C. and Pardo, T. A. S. (2025). Propbanks e representações semânticas: o que temos, o que queremos e o que podemos. *LinguaMÁTICA*, 17(2):1–29.

Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Silva, J., and Aluísio, S. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks.

In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, pages 122–131.

Hartmann, N. S., Duran, M. S., and Aluísio, S. M. (2016). Automatic semantic role labeling on non-revised syntactic trees of journalistic texts. In *International Conference on Computational Processing of the Portuguese Language*, pages 202–212.

Heinecke, J. (2023). metamorphosed, a graphical editor for abstract meaning representation. In *Proceedings of the 19th Joint ACL-ISO Workshop on Interoperable Semantics (ISA-19)*, pages 27–32, Nancy, France. Association for Computational Linguistics.

Hermjakob, U. (2013). Amr editor: A tool to build abstract meaning representations. `https://amr.isi.edu/papers/amr-editor-ulf2013a.pdf`. Accessed: 25 Aug 2024.

Ho, S. H. (2025). Evaluation of finetuned llms in amr parsing. *arXiv preprint arXiv:2508.05028*. 27 pages, 32 figures.

Inácio, M. L., Cabezudo, M. A. S., Ramisch, R., Di-Felippo, A., and Pardo, T. A. S. (2023). The amr-pt corpus and the semantic annotation of challenging sentences from journalistic and opinion texts. *DELTA: Documentação de Estudos em Linguística Teórica e Aplicada*, 39(3):1–31.

Jurafsky, D. and Martin, J. H. (2026). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*. 3rd edition. Online manuscript released January 6, 2026.

Li, Y. and Fowlie, M. (2025). Gpt makes a poor amr parser. *Journal for Language Technology and Computational Linguistics*, 38(2):43–76.

Lopes, L. and Pardo, T. (2024). Towards portparser - a highly accurate parsing system for Brazilian Portuguese following the Universal Dependencies framework. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 401–410, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Mansouri, B. (2025). Survey of abstract meaning representation: Then, now, future.

Meyes, R., Lu, M., de Puiseau, C. W., and Meisen, T. (2019). Ablation studies in artificial neural networks. *CoRR*, abs/1901.08644.

Oliveira, S., Loureiro, D., and Jorge, A. (2021). Improving portuguese semantic role labeling with transformers and transfer learning. In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–9.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Sadeddine, Z., Opitz, J., and Suchanek, F. (2024). A survey of meaning representations – from theory to practical utility. In Duh, K., Gomez, H., and Bethard, S., editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2877–2892, Mexico City, Mexico. Association for Computational Linguistics.

Sanches Duran, M. and Aluísio, S. (2015). Automatic generation of a lexical resource to support semantic role labeling in Portuguese. In Palmer, M., Boleda, G., and Rosso, P., editors, *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 216–221, Denver, Colorado. Association for Computational Linguistics.

Sobrevilla Cabezudo, M. A. and Pardo, T. (2019). Towards a general Abstract Meaning Representation corpus for Brazilian Portuguese. In Friedrich, A., Zeyrek, D., and Hoek, J., editors, *Proceedings of the 13th Linguistic Annotation Workshop*, pages 236–244, Florence, Italy. Association for Computational Linguistics.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In *Proceedings of the 9th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 403–417, Cham. Springer.

Straka, M., Straková, J., and Gamba, F. (2024). ÚFAL LatinPipe at EvaLatin 2024: Morphosyntactic analysis of Latin. In Sprugnoli, R. and Passarotti, M., editors, *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)*, pages 207–214, Torino, Italia. ELRA and ICCL.

Vasylenko, P., Huguet Cabot, P. L., Martínez Lorenzo, A. C., and Navigli, R. (2023). Incorporating graph information in transformer-based AMR parsing. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1995–2011, Toronto, Canada. Association for Computational Linguistics.

Wang, C., Xue, N., and Pradhan, S. (2015). Boosting transition-based AMR parsing with refined actions and auxiliary analyzers. In Zong, C. and Strube, M., editors, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 857–862, Beijing, China. ACL.

Wein, S. and Opitz, J. (2024). A survey of AMR applications. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6856–6875, Miami, Florida, USA. Association for Computational Linguistics.

Zhou, J., Naseem, T., Fernandez Astudillo, R., Lee, Y.-S., Florian, R., and Roukos, S. (2021). Structure-aware fine-tuning of sequence-to-sequence transformers for transition-based AMR parsing. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6279–6290, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Comparative Framework for Diachronic Lexical Semantic Change in Brazilian Portuguese Political Discourse

João Victor M. Correa[1] **, Oto Araújo Vale**[2] **, Renato Moraes Silva**[1]

[1]Interinstitutional Center for Computational Linguistics (NILC), Institute of Mathematics and Computer Science, University of São Paulo, São Carlos, Brazil

[2]Federal University of São Carlos (UFSCar), São Carlos, São Paulo, Brazil

victormcorrea@usp.br, otovale@ufscar.br, renatoms@icmc.usp.br

Abstract. *Shifts in political discourse often reflect societal changes, but quantifying how lexical usage and semantic representations vary over time remains a challenge for computational linguistics. To compare approaches under different computational constraints, we propose a cost-aware, model- and language-agnostic framework to compare lexical and semantic change signals. We apply it to Brazilian Portuguese political discourse using a shared panel of lemmas combining drift candidates, stable controls, and theory-driven seeds. We compare three families of diachronic drift signals: a TF-IDF lexical-salience baseline, aligned slice-specific Word2Vec models, and a contextual BERT approach. Results indicate disagreement between cheaper baselines, while contextual BERT remains weakly aligned with each method. A filtered contextual panel reduces stable-control leakage and highlights drift candidates such as "bloqueio" and "salário". Rather than treating any detector as ground truth, this study presents an exploratory comparison of complementary drift sensitivities, interpretability, and computational cost in Brazilian Portuguese political discourse.*

1. Introduction

Political language carries historical change in its lexical choices, as shifts in ideological agendas, coalition structures, and policy framing leave traces in word usage over time. Analyzing these large-scale patterns by hand is impractical. Diachronic natural language processing (NLP) methods can capture some of the movement at scale, but different modeling choices foreground different aspects of change [Hamilton et al. 2016, Kutuzov et al. 2018, Tahmasebi et al. 2021]. In political corpora the picture is especially complex, since a term may drift because of a new policy referent, a rhetorical reframing, a shift in collocational profile, or a reorganization of discourse coalitions. No single mechanism accounts for all of them.

This complexity carries over to evaluation, where shared-task benchmarks have improved comparability in lexical semantic change detection [Schlechtweg et al. 2020], but curated external labels remain scarce, especially outside the few languages and corpora covered by standard tasks. Moreover, distributional pipelines can be affected by temporal artifacts, alignment noise, and other confounds when outputs are read too literally [Dubossarsky et al. 2019]. Without externally validated ground truth, an important issue is what different detectors capture and where their rankings agree or diverge on the same political timeline. This question is timely for Brazilian Portuguese because, despite the recent availability of large-scale political corpora such as BrPoliCorpus [Lima-Lopes 2025], comparative studies benchmarking multiple drift detection families on the same Portuguese political corpus remain scarce.

To address this gap, we propose a cost-aware comparative framework for diachronic lexical and semantic change analysis. The framework stages detection methods into lexical-statistical, static-embedding, and contextual tiers, then compares their outputs on a shared

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 100–112
October 19-22, 2026 ©2026 Association for Computational Linguistics

panel rather than letting each method define its own evidence base. We instantiate it on the floor subset of BrPoliCorpus and evaluate TF-IDF, aligned Word2Vec, and BERTimbau Large on the same set of lemmas. The paper addresses the following research questions:

1. How strongly do the drift detection methods agree on the same lemmas?
2. Which lemmas identified as drift candidates are method-specific?
3. Does the higher-cost contextual stage of the proposed framework add enough interpretive value to justify its computational expense?

2. Related Work

Distributional semantic change detection evolved from co-occurrence comparisons to aligned static embeddings and then to contextual representations [Kutuzov et al. 2018, Tahmasebi et al. 2021]. Among static approaches, slice-specific Word2Vec models aligned with orthogonal Procrustes can recover known historical shifts and reveal statistical regularities of change [Hamilton et al. 2016]. Such embeddings remain attractive because they are efficient and produce interpretable neighborhood movements, but compress all word usages into a single vector, blurring cases where older and newer meanings coexist. Moreover, evaluation remains a central challenge. Standardized shared tasks have improved comparability [Schlechtweg et al. 2020], yet many real-world studies still operate without external semantic ground truth, and apparent change can emerge from temporal referencing, alignment noise, or genre shifts rather than from semantic drift alone [Dubossarsky et al. 2019]. These concerns motivate the stable controls and agreement analysis proposed in this study.

Contextual models provide a complementary perspective. BERT-style representations encode token-level usages and can surface variation that static spaces smooth over [Giulianelli et al. 2020, Devlin et al. 2019]. Scalable contextual methods that improve interpretability have also been proposed [Montariol et al. 2021], and pre-trained Portuguese checkpoints such as BERTimbau [Souza et al. 2020] have made contextual analysis feasible for Portuguese NLP.

Diachronic analysis of political language has gained traction through parliamentary corpora. The study of [Hofmann et al. 2020] compared lexical usage in Austrian political discourse across time, but comparable studies of Brazilian Portuguese political speech are still scarce. Available Portuguese diachronic resources have focused primarily on historical and literary material rather than political discourse [Osório and Cardoso 2025]. BrPoliCorpus [Lima-Lopes 2025] now provides the scale needed for controlled comparative work in this domain. This literature motivates a comparison in which no single detector is treated as the ground truth. Our framework makes the comparison explicit because cheaper methods screen broadly, contextual modeling is used selectively, and agreement diagnostics show where the resulting rankings converge or diverge.

3. Comparative Drift Framework

We first describe the framework in general terms, independent of any particular corpus or model. Section 4 then instantiates it with BrPoliCorpus, TF-IDF, Word2Vec, and BERTimbau Large. Figure 1 summarizes the overall flow.

The framework assumes a diachronic corpus partitioned into temporal slices, where each slice must contain enough documents and tokens to support stable frequency estimates and distributional model training. A preprocessing step then normalizes surface forms to a shared lemma representation so that all three tiers operate on the same vocabulary. Not every word form is suitable for drift scoring, so the framework applies three joint filters before any method runs: a minimum per-slice frequency to guarantee sufficient token mass for embedding training, a document-dispersion threshold to prevent a single source

from inflating apparent salience, and a minimum slice-presence rate to restrict attention to genuinely diachronic terms rather than episodic ones. Drift and stable-control candidates may be further restricted to dominant part-of-speech classes.

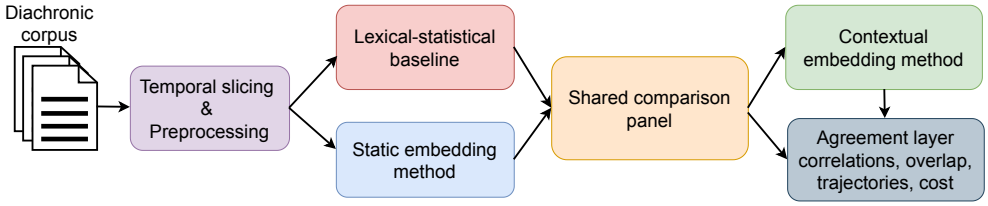


Figure 1. Staged framework for comparing lexical and semantic change signals. Two cheaper tiers screen the vocabulary, whose candidates are combined in a shared panel before the contextual tier and agreement analysis.

The framework then organizes comparative change analysis into three computational tiers arranged by increasing cost. A **lexical-statistical baseline** captures shifts in lexical salience with no model training. A **static-embedding method** aligns one vector space per temporal slice and measures neighborhood displacement. A **contextual-embedding method** represents sampled token usages with a pretrained language model, which can surface usage-level variation that cheaper type-level methods smooth over, but at much higher computational cost.

The key design principle is *staged cost awareness*. The two cheaper tiers first score the full eligible vocabulary, using their own representations to estimate diachronic drift for each lemma. Their top candidates feed into the expensive contextual stage, avoiding costly inference over every eligible lemma. This ordering follows the practical distinction between broad type-based discovery and costlier token-based analysis in lexical semantic change work [Kurtyigit et al. 2021]. Cross-method comparison then requires a fixed evaluation set, so no method is judged only on the terms it would naturally select. The framework therefore prescribes a **shared comparison panel** assembled from four sources: (i) high-drift candidates nominated by the lexical-statistical baseline; (ii) high-drift candidates nominated by the static-embedding method; (iii) low-drift stable controls (lemmas drawn from the low-drift tail of each cheaper tier), which act as negative examples to help detect spurious drift signals rather than only highlighting positive cases [Dubossarsky et al. 2017]; and (iv) theory-driven seed terms drawn from domain expertise or prior work. Because all three tiers score the same candidate set, any disagreement reflects genuine differences in what each method measures rather than divergent vocabularies.

Finally, the framework adds an **agreement analysis layer**, which compares the rankings produced after all tiers score the same shared panel. This layer comprises four complementary analyses. First, pairwise rank correlations measure agreement across the full ranking. Second, top-k overlap curves quantify agreement among high-ranked lemmas as k grows. Third, stable-control leakage rates measure whether stable controls remain low ranked by identifying controls that rise into candidate regions, exposing method-side noise, as emphasized by control-based analyses of semantic change models [Dubossarsky et al. 2017]. Fourth, after leaked controls are removed, the filtered contextual ranking identifies the contextual candidates most suitable for qualitative inspection.

The framework is model- and language-agnostic, allowing any combination of methods respecting the three-tier cost ordering; the next section details our implementation.

4. Methodology

We instantiate the framework on Brazilian Portuguese parliamentary discourse, specifying the corpus, preprocessing choices, and the model used in each tier.

4.1. Corpus and Preprocessing

More diachronic Portuguese resources are now available [Osório and Cardoso 2025], ranging from literary collections such as the Tycho Brahe Parsed Corpus [Galves and Faria 2017] to broad repositories such as the Corpus do Português [Davies and Ferreira 2006]. However, these resources target broad historical periods rather than domain-specific political analysis with continuous yearly slices. BrPoliCorpus [Lima-Lopes 2025] addresses this gap by providing large-scale parliamentary data for Brazilian Portuguese.

We performed experiments on its subset containing Brazilian Chamber floor speeches from 2000 to 2023, organized into 24 yearly slices. This floor subset offers the largest contiguous yearly coverage (428,366 speeches across 24 slices), a uniform parliamentary register, and exact date metadata, making it the most suitable segment for a year-by-year comparative drift study in the political domain. Here, diachronic analysis denotes comparison across time-ordered corpus slices rather than a fixed minimum historical span. Temporal granularity varies with the research question and available data, with prior studies using monthly or yearly slices for socio-cultural change and longer bins for long-term linguistic change [Kutuzov et al. 2018]. Our 24 yearly slices yield 23 adjacent-year transitions and support analyses of annual change in political discourse.

As described in Section 3, not all tokens are eligible for drift scoring. Raw speeches are processed using `spaCy` (`pt_core_news_sm`) [Honnibal et al. 2020]. We retain nouns, verbs, and adjectives, exclude proper nouns, remove stopwords, punctuation, and numeric tokens, and keep only tokens with at least two characters. All text is lowercased while preserving accents. Each token is then mapped to its lemma form, producing the textual representation used by all three tiers of the proposed framework (Figure 1). The experimental environment used spaCy 3.8.11. The parser and named-entity recognizer were disabled, while the model's morphologizer and lemmatizer remained active. A fixed correction list repaired selected lemma errors and malformed Portuguese infinitives. After preprocessing, the corpus retains 428,366 speeches and 63,036,642 tokens across 225,484 unique lemmas. Table 1 summarizes the corpus totals and the quartile structure of yearly variation.

A lemma enters the scored vocabulary only if it appears at least 50 times per slice, in at least 5 documents per slice, and in at least 80% of the 24 slices. These thresholds were fixed as practical corpus-stability defaults rather than tuned to optimize any particular method, and may vary when applied to corpora with different size, genre composition, or temporal density.

The frequency floor ensures sufficient token mass for reliable embedding training in each yearly slice; the document-dispersion requirement prevents a single speech from inflating a lemma's apparent salience; and the slice-presence constraint restricts attention to lemmas that are genuinely diachronic rather than episodic. Drift and stable-control candidates are further restricted to nouns and adjectives, and a small set of procedural high-frequency terms is excluded. These filters reduce the scored vocabulary to 3,221 lemmas for the two cheaper tiers of the framework (Section 3).

Yearly document counts range from 3,041 to 26,868 (median 16,752) and retained token volume from 630K to 4.2M (median 2.6M), which is why the frequency and slice-presence thresholds matter in practice.

4.2. Drift Detection Tiers

With the eligible vocabulary defined, we now describe the well-established model used in each tier of the framework. The lexical-statistical tier uses a TF-IDF lexical-salience baseline, the static-embedding stage uses slice-aligned Word2Vec with Orthogonal Procrustes, and the contextual stage uses BERTimbau Large. These individual techniques are standard

Table 1. Corpus summary for the BrPoliCorpus floor subset. Yearly slice statistics are reported as the 25th percentile (Q1), median, and 75th percentile (Q3) across the 24 yearly files; token and character counts are shown in millions.

Period	Slices	Measure	Corpus total	Yearly slices		
				Q1	Median	Q3
2000–2023	24	Speeches	428,366	14,648	16,752	21,987
		Tokens	63.0M	2.00	2.62	3.45
		Characters	538.5M	16.9	22.3	29.7
		Lemmas	225,484	39,456	50,269	58,637

in the literature [Hamilton et al. 2016, Giulianelli et al. 2020]; the contribution here lies in combining them inside the staged comparison.

For the **lexical-statistical tier**, we treat each yearly slice as one document in the corpus-level TF-IDF computation. Term frequency normalizes a lemma's count by the slice's total token count, while smoothed inverse-document frequency is computed across the 24 slices, so that a lemma's document frequency corresponds to the number of slices in which it appears. For each eligible lemma, we then read its TF-IDF weight from each slice, yielding a 24-entry temporal trajectory. Drift is defined as the mean absolute change in weight between consecutive slices. This provides a continuous ranking of shifts in lexical salience, reflecting how central a term becomes to each year's agenda, without relying on distributional neighborhoods. For lemma w and year t, with $N = 24$ yearly slices, the smoothed weight and change score are $\mathrm{idf}(w) = \log[(1 + N)/(1 + \mathrm{df}(w))] + 1$, $x_{w,t} = \mathrm{tf}(w, t)\,\mathrm{idf}(w)$, and $s_{\text{TF-IDF}}(w) = \frac{1}{N-1}\sum_{t=2}^{N} |x_{w,t} - x_{w,t-1}|$. The absolute difference prevents increases and decreases from cancelling. Each lemma-year pair has one scalar weight rather than a word vector, so cosine distance would not provide a meaningful comparison for this representation.

For the **static-embedding tier**, we train independent **Word2Vec** (skip-gram) models per yearly slice (300 dimensions, window 5, 5 negative samples, minimum count 5, 5 epochs, and 2 replicates with different seeds) and align them into a common space using Orthogonal Procrustes over the top 20,000 anchor words [Hamilton et al. 2016]. Drift is computed as the mean cosine distance between a lemma's aligned vectors in consecutive slices, averaged over replicates. This captures persistent neighborhood reorganization.

For the **contextual tier**, we embed sampled in-context occurrences using a Portuguese BERT checkpoint (**BERTimbau Large** [Souza et al. 2020]), extracting hidden states from layers -4 (an upper-mid layer) and -1 (the final layer). The choice of these two layers follows systematic studies of contextual lexical-change models, which report that semantic-change signals are often strongest in upper-mid layers, while the final layer is more exposed to surface-form information that can bias change scores [Laicher et al. 2021, Periti and Tahmasebi 2024]. We therefore use layer -4 as a robustness comparison and layer -1 as the primary ranking. For each lemma and slice, up to 64 occurrences are sampled using up to 48 pretokenized words on each side of the target. The contextual drift score is computed as the mean cosine distance between slice-level centroid embeddings across consecutive slices.

4.3. Shared Comparison Panel and Agreement Layer

Following the framework design (Section 3), we build one shared panel before running the contextual stage. The panel contains 55 lemmas: 15 Word2Vec-led drift candidates, 15 TF-IDF-led drift candidates, 20 stable controls, and 5 theory seeds. The values of this composition are fixed a priori as a pragmatic design choice considering three practical

concerns: coverage of each method's top region, enough negative examples for the leakage diagnostics, and BERT inference time over 24 yearly slices. The method-led candidates are the highest-ranked nouns or adjectives from each cheaper tier after the eligibility filters described in Section 4.1. Limiting each tier to 15 items keeps the panel within the informative top region without inflating contextual cost, since each additional lemma requires sampled contexts across all slices. Including 20 stable controls terms ensures that leakage diagnostics have enough negative examples to detect spurious drift signals. The 5 theory seeds are *democracia*, *corrupção*, *reforma*, *economia*, and *liberdade*. We selected them by design as broad political vocabulary anchors, based on domain knowledge rather than a formal selection criterion, and used them as an interpretive reference rather than as an additional set of drift candidates.

All three methods score this same panel. The panel contains 33 nouns and 22 adjectives, which are ranked jointly rather than in separate part-of-speech lists. For display, each method is reranked within the fixed panel and rank r is converted to $p = 1 - (r-1)/(N-1)$ with $N = 55$. Thus, $p = 1$ denotes the highest score and $p = 0$ the lowest. This is an ordinal display scale, not a probability or a comparable effect size. The agreement layer then reports pairwise Spearman rank correlations, top-k overlap curves for $k = 1, \ldots, 15$, stable-control leakage, and a filtered contextual ranking for qualitative inspection. To assess whether methods meaningfully separate drift candidates from stable controls, we also compare their within-panel rank-percentile distributions using one-sided Mann-Whitney tests, where the alternative hypothesis is that drift candidates receive systematically higher ranks.

5. Results

Figure 2 reports the pairwise Spearman rank-correlation diagnostic from the agreement analysis layer. Each point is one of the 55 shared-panel lemmas, and each panel compares how two methods score the same lemma. The four panels compare BERT vs. Word2Vec, BERT vs. TF-IDF, Word2Vec vs. TF-IDF, and BERT layers -4 vs. -1.

The two lighter baselines disagree strongly even after they are constrained to the same candidate universe. As shown in Figure 2, the Word2Vec vs. TF-IDF panel displays a clear negative trend; lemmas ranked highly by one baseline tend to be ranked low by the other. Quantitatively, the two methods reach Spearman $\rho = -0.54$ (p-value $= 2.1 \times 10^{-5}$), showing that they order the shared panel differently.

Figure 2 also shows that the contextual tier partly mediates this split without collapsing onto either baseline. In the BERT vs. Word2Vec and BERT vs. TF-IDF panels, contextual BERT displays weak but positive associations with both methods. On the primary layer (-1), BERT reaches $\rho = 0.21$ (p-value $= 0.128$) with Word2Vec and $\rho = 0.12$ (p-value $= 0.365$) with TF-IDF.

The BERT layer comparison gives a separate robustness check. Layers -4 and -1 agree closely (Spearman $\rho = 0.858$, p-value $< 10^{-15}$), so the contextual conclusions are not driven by a single extraction depth.

Figure 3 reports the remaining diagnostics of the agreement layer. Panel A shows top-k overlap across methods, while Panels B–D use standard boxplots to compare the within-panel rank-percentile distributions of drift-candidate buckets and stable controls for Word2Vec, TF-IDF, and BERT. In the primary BERT layer (-1), the top-15 overlap is 7 terms with Word2Vec and 6 terms with TF-IDF, whereas the baseline-to-baseline overlap remains zero. Thus, contextual BERT recovers candidates favored by both cheaper tiers while preserving an intermediate ranking structure of its own.

The boxplots in Panels B–D further evaluate whether each method separates nominated drift terms from stable controls. Word2Vec and TF-IDF rank drift candidates signifi-

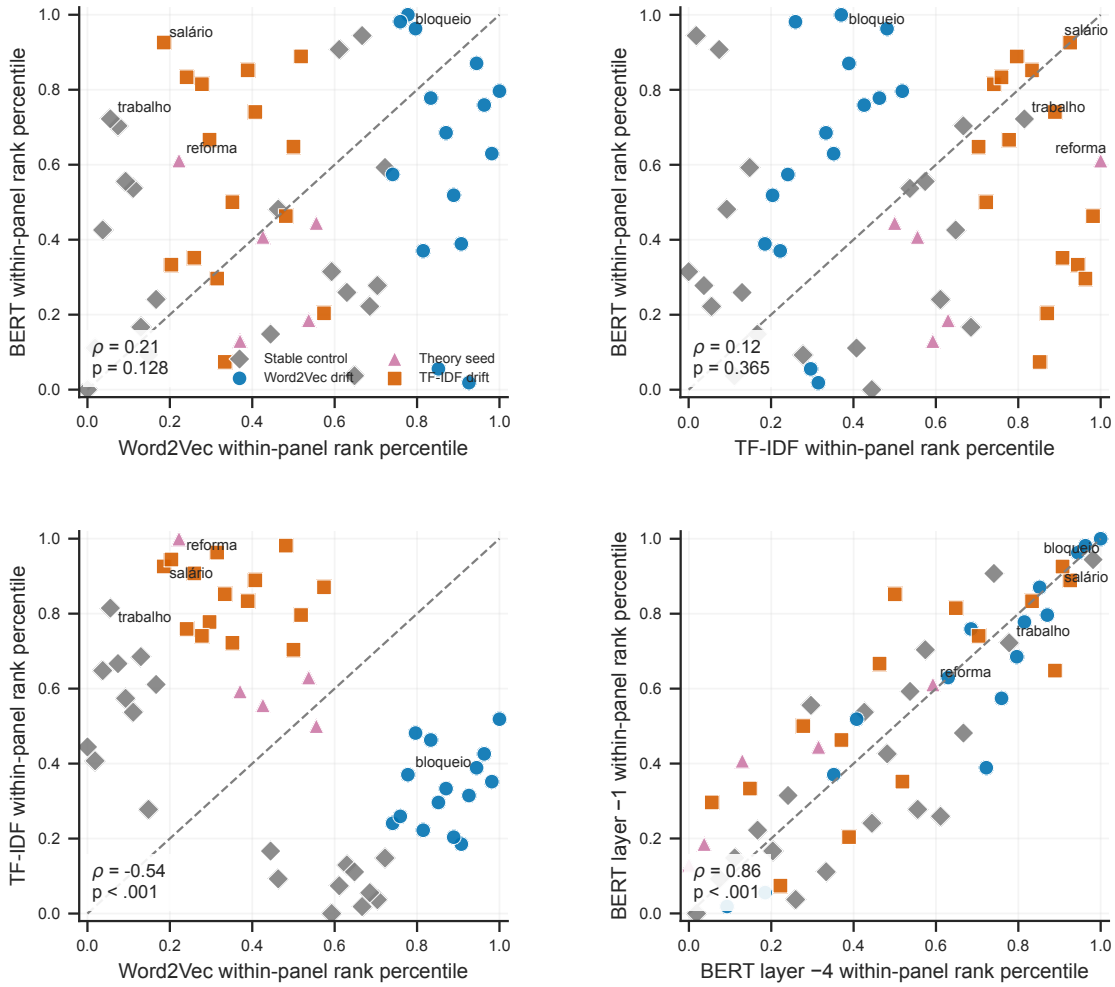


Figure 2. Pairwise within-panel rank agreement from the agreement layer on the 55-lemma shared panel: BERT vs. Word2Vec (top left), BERT vs. TF-IDF (top right), Word2Vec vs. TF-IDF (bottom left), and BERT layers -4 vs. -1 (bottom right). Points are lemmas; annotations report Spearman ρ and two-sided p-values. Within-panel percentiles range from 1 for the highest change score to 0 for the lowest.

cantly above stable controls (p-value < 0.001 in both cases, one-sided Mann–Whitney), showing clean bucket separation. BERT also separates the groups, but with weaker margins (p-value $= 0.008$). For that reason, the contextual ranking is most useful as a selective interpretive layer rather than as an unrestricted final candidate list. After filtering leaked controls, the remaining contextual candidates provide informative arbitration among terms proposed by the cheaper methods.

As a robustness check, we recomputed contextual scores using average pairwise distance (APD) [Giulianelli et al. 2020] instead of the centroid-based metric. On the primary layer, APD and centroid rankings correlate at Spearman $\rho = 0.79$ (p-value $< 10^{-12}$), with 7 of 15 top terms shared. The two metrics foreground different properties. APD promotes high-frequency polysemous terms whose embedding clouds are broad (e.g. *trabalho*, *público*), while the centroid metric favours terms with coherent directional shifts (*bloqueio*, *salário*). Notably, the centroid metric separates method-nominated drift candidates from stable controls more cleanly than APD (p-value $= 0.008$ vs. p-value $=$

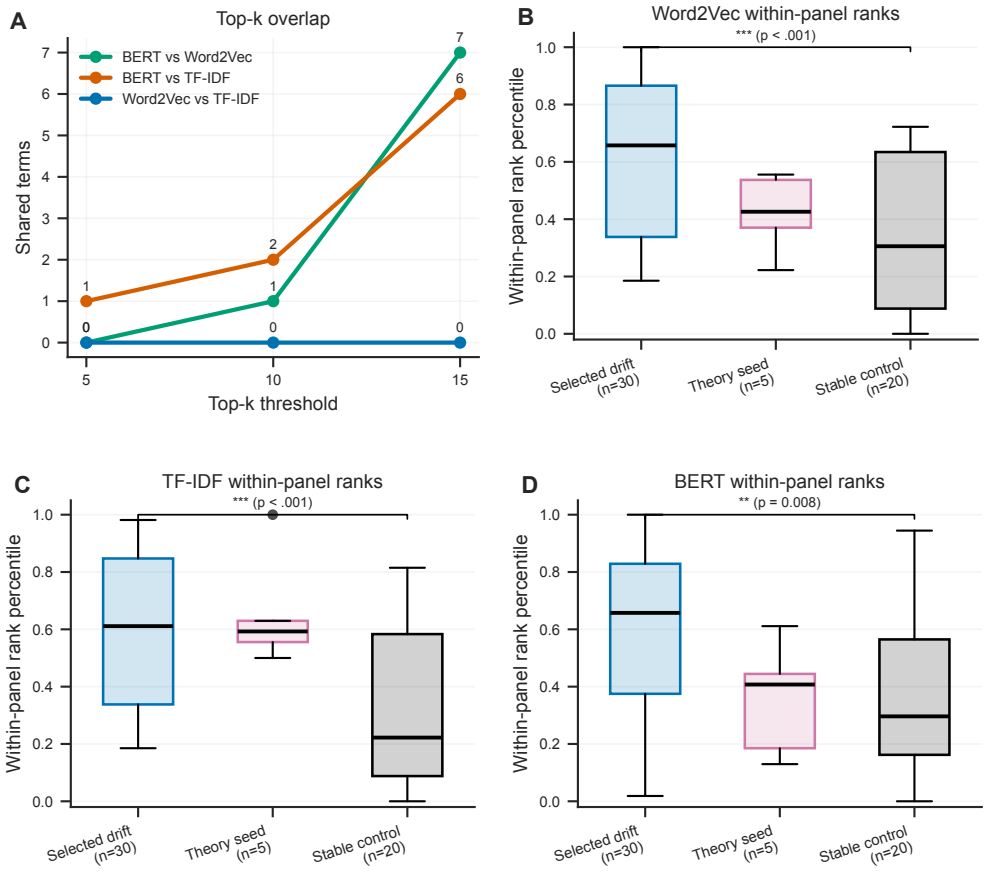


Figure 3. Overlap and rank-distribution summaries on the shared panel. Panel A shows top-k overlap between methods. Panels B–D show boxplots of within-panel rank-percentile distributions by bucket for Word2Vec, TF-IDF, and BERT. Brackets indicate one-sided Mann–Whitney tests for drift terms vs. stable controls.

0.081, Mann–Whitney).

To connect these rankings with observed usage, Table 2 presents paired early and late occurrences for one candidate nominated by each cheaper method and for a stable control. The excerpts support qualitative inspection, but they are not annotated semantic-change labels.

Table 2. Illustrative corpus occurrences for two method-nominated drift candidates and one stable control. BERT ranks refer to the primary layer (-1) within the 55-lemma shared panel. The symbol "…" indicates omitted text.

Term and selection	Early occurrence (year 2000)	Late occurrence (year 2023)
bloqueio Word2Vec-led; BERT 1/55	"… o terrorismo é um apêndice, uma alternativa do bloqueio para golpear a nação cubana …"	"… a falta de água, comida e energia elétrica, e o bloqueio para crianças na Faixa de Gaza."
salário TF–IDF-led; BERT 5/55	"… o Brasil vive, no momento, a grande discussão sobre o salário mínimo."	"… reajustar o valor do salário mínimo acima da inflação; … recuperar o poder de compra …"
trabalho Stable control; BERT 16/55	"… o desemprego oculto pelo trabalho precário …"	"… ilustre amigo que faz muito bem o seu trabalho."

The examples illustrate distinct behaviors captured by the framework and the need to distinguish drift signals from confirmed semantic change. *Bloqueio*, a Word2Vec-led candidate ranked 1/55 by BERT, exhibits a change in referent and geopolitical frame across occurrences, consistent with contextual or neighborhood movement but not sufficient to establish a conventionalized new sense. *Salário*, a TF–IDF-led candidate ranked 5/55 by BERT, remains associated with minimum-wage policy in both excerpts, suggesting increased political salience rather than clear semantic change. Finally, *trabalho*, a stable control ranked 16/55 by BERT, illustrates ordinary contextual variation and why drift scores and selected excerpts should not be interpreted as semantic-change gold labels.

6. Discussion

To examine how the methods differ at the level of individual terms, Figure 4 plots standardized temporal trajectories for four representative lemmas. The selected cases cover four informative situations: a Word2Vec-led term with strong contextual support (*bloqueio*), a TF-IDF-led term with strong contextual support (*salário*), a theory seed whose contextual rank rises despite modest static drift (*reforma*), and a stable-control diagnostic term (*trabalho*) that reveals contextual leakage.

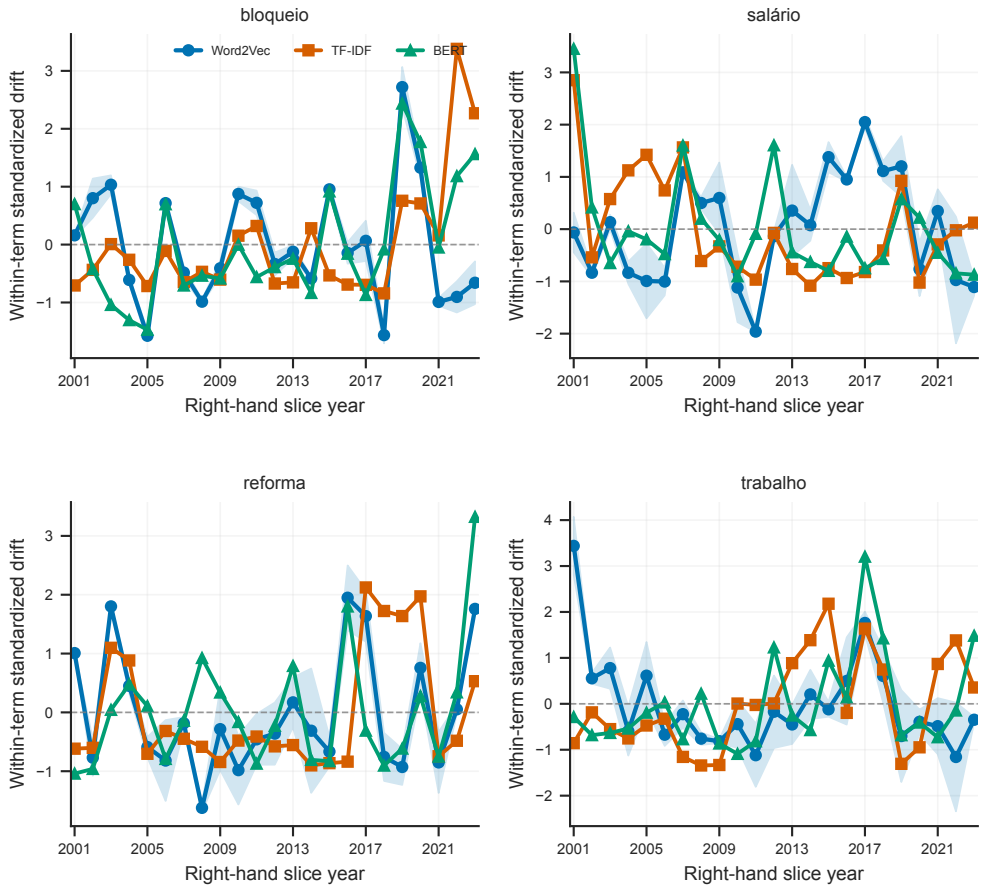


Figure 4. Representative term trajectories for *bloqueio* (top left), *salário* (top right), *reforma* (bottom left), and *trabalho* (bottom right). Scores are standardized within term and method to emphasize temporal shape rather than absolute scale. Word2Vec ribbons denote replicate variability.

The disagreement itself is substantively informative, because each method foregrounds a different facet of lexical change. TF-IDF is most sensitive to lexical-profile

reweightings, so a term whose salience rises in yearly agendas, policy disputes, or formulaic framing can register as drifting even if its local semantic neighborhood remains comparatively stable. Word2Vec, by contrast, captures persistent neighborhood reorganization across slices, highlighting terms whose distributional associations reorganize over time. Contextual BERT occupies an intermediate position and can support either a TF-IDF-led or a Word2Vec-led candidate through its token-level representations, although the stable-control leakage analysis shows that greater context sensitivity does not automatically produce a cleaner unrestricted ranking.

The paired occurrences in Table 2 help constrain the interpretation of these rankings. *Salário* behaves like a lexical-profile case with strong contextual support, but both excerpts retain the minimum-wage sense. Its ranking therefore more directly supports changing political salience than semantic change. In the selected *bloqueio* excerpts, the geopolitical referent shifts from Cuba to Gaza while the core sense of a blockade remains stable. Its ranking is therefore more consistent with contextual and distributional reorganization than with clear sense innovation. *Reforma* illustrates a complementary pattern; it is only modestly ranked by the cheaper tiers, but it is ranked higher by the contextual model, a rank difference that does not establish greater accuracy in detecting semantic change. Finally, the stable-control excerpts for *trabalho* illustrate ordinary polysemy, while its BERT rank shows that context sensitivity can still promote a term expected to remain stable, motivating the leakage diagnostics in the agreement layer.

Table 3. Method scope and computational cost (estimated relative to TF-IDF).

Method	Signal focus	Scope	Relative cost
TF-IDF	Adjacent-slice salience shift	3,221 lemmas	seconds ($1\times$)
Word2Vec	Aligned neighborhood displacement	3,221 lemmas	minutes ($\sim 56\times$)
BERT	Usage divergence from sampled contexts	55 lemmas / 82,461 occurrences	hours ($\sim 600\times$)

Table 3 quantifies the cost gap observed in our single-machine experimental setting. Given the modest rank correlations reported in Section 5, the current evidence does not support replacing the cheaper methods with contextual models. The pattern supports the staged design; cheaper methods screen the vocabulary broadly, while contextual modeling is applied selectively when disagreement is substantively interesting.

7. Conclusion

This paper introduced a shared, cost-aware comparative framework for contrasting three families of lexical and semantic change signals in Brazilian Portuguese political discourse. The research questions posed in the introduction receive three answers from the evidence. First, regarding the strength of inter-method agreement, TF-IDF and Word2Vec disagree strongly on the shared panel and contextual BERT remains only weakly correlated with either baseline. Second, regarding method-specific candidates, the two cheaper methods identify largely non-overlapping sets of drift terms. BERT partially bridges this split by placing candidates from both camps among its top-ranked terms, but it also retains some stable controls in high positions, so its raw ranking requires stable-control filtering before qualitative interpretation.

Third, regarding the cost-benefit of the contextual stage, BERT provides useful adjudication when inter-method disagreement is substantively interesting, as illustrated by terms such as *bloqueio*, *salário*, and *reforma*. At roughly $600\times$ the cost of TF-IDF, though, the current evidence favors using it selectively rather than as a replacement for the cheaper methods. Because the comparative framework specifies roles and interfaces, it can be instantiated with alternative static or contextual embeddings.

8. Limitations

The primary limitation is the lack of external semantic ground truth for the BrPoliCorpus floor subset. Therefore, agreement and disagreement are analyzed as comparative evidence rather than as proof that one detector is correct. A second, methodological limitation is that the raw contextual ranking still admits some stable controls near the top, even though the filtered contextual list is useful for interpretation; contextual sensitivity does not automatically yield a cleaner unrestricted candidate list. The choice of contextual aggregation strategy also merits discussion. We report centroid-based drift as the primary contextual metric and verify it against APD [Giulianelli et al. 2020, Periti and Tahmasebi 2024] (Section 5); the strong correlation and the centroid metric's tighter bucket separation support this choice, though APD captures complementary embedding-cloud variance. Sense-based clustering methods [Montariol et al. 2021] were not attempted here. The TF-IDF tier also treats each annual slice as one aggregate document. Token-normalized term frequency reduces differences in yearly corpus size but does not control speaker composition, verbosity, or topical breadth. Retrieval research on document-length normalization illustrates this broader concern [Lipani et al. 2015]; a speech-level or scope-aware count baseline would use a different document unit and require separate evaluation. Finally, yearly slicing is appropriate for a first comparative baseline but compresses within-year shocks, campaign cycles, and short-lived institutional disputes that may matter for political language.

Future work could address the evaluation gap through external annotation or noisy supervision from labeled parliamentary datasets, though such resources still require careful deduplication before serving as reliable comparison targets. On the modeling side, adopting task-specialized models such as XL-LEXEME [Cassotti et al. 2023], exploring newer divergence metrics [Goworek and Dubossarsky 2026], and applying orthographic-bias mitigation through subword masking [Laicher et al. 2021] would bring the contextual stage closer to current best practices. A symbolic companion analysis using lexical complexity or readability features could also help separate lexical-semantic drift from rhetorical or stylistic change.

9. Ethics Statement

The study analyzes public institutional political speech and reports only aggregate corpus statistics, ranked lemmas, and usage-level comparative signals. We do not profile individual speakers or infer private attributes. Even so, political corpora encode unequal representation and occasionally harmful rhetoric, so drift scores should not be read as normative judgments, evidence of speaker intent, or standalone claims about social meaning. All outputs are exploratory comparative signals whose interpretation depends on corpus composition, preprocessing choices, and the distinction between lexical salience, contextual variation, and broader discourse change.

10. Generative AI Use Disclosure

Generative AI tools were used to suggest alternatives for rewriting some sentences in a concise way and to correct spelling. All suggestions were manually verified by the authors.

11. Code Availability

The source code for the comparative drift framework is available at: `https://github.com/nilc-nlp/brpolicorpus-semantic-drift`.

12. Acknowledgements

We gratefully acknowledge the support provided by the São Paulo Research Foundation (FAPESP; grant No. 2024/17834-6).

References

Cassotti, P., Siciliani, L., DeGemmis, M., Semeraro, G., and Basile, P. (2023). XL-LEXEME: WiC pretrained model for cross-lingual LEXical sEMantic changE. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1577–1585, Toronto, Canada. Association for Computational Linguistics.

Davies, M. and Ferreira, M. J. (2006). Corpus do português. Available at `http://www.corpusdoportugues.org`. Accessed: May 15, 2026.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dubossarsky, H., Grossman, E., and Weinshall, D. (2017). Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145, Copenhagen, Denmark. Association for Computational Linguistics.

Dubossarsky, H., Hengchen, S., Tahmasebi, N., and Schlechtweg, D. (2019). Time-out: Temporal referencing for robust modeling of lexical semantic change. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 457–470. Association for Computational Linguistics.

Galves, C. and Faria, P. (2017). Tycho Brahe parsed corpus of historical Portuguese.

Giulianelli, M., Del Tredici, M., and Fernández, R. (2020). Analysing lexical semantic change with contextualised word representations. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

Goworek, R. and Dubossarsky, H. (2026). Rethinking metrics for lexical semantic change detection. In Tahmasebi, N., Cassotti, P., Montariol, S., Kutuzov, A., Huebscher, N., Spaziani, E., and Baes, N., editors, *The Proceedings for the 6th International Workshop on Computational Approaches to Language Change (LChange'26)*, pages 147–161, Rabat, Morocco. Association for Computational Linguistics.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In Erk, K. and Smith, N. A., editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Hofmann, K., Marakasova, A., Baumann, A., Neidhardt, J., and Wissik, T. (2020). Comparing lexical usage in political discourse across diachronic corpora. In *Proceedings of the Second ParlaCLARIN Workshop*, pages 58–65, Marseille, France. European Language Resources Association.

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). spacy: Industrial-strength natural language processing in python. DOI: 10.5281/zenodo.1212303.

Kurtyigit, S., Park, M., Schlechtweg, D., Kuhn, J., and Schulte im Walde, S. (2021). Lexical semantic change discovery. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference*

on Natural Language Processing (Volume 1: Long Papers), pages 6971–6981, Online. Association for Computational Linguistics.

Kutuzov, A., Øvrelid, L., Szymanski, T., and Velldal, E. (2018). Diachronic word embeddings and semantic shifts: a survey. In Bender, E. M., Derczynski, L., and Isabelle, P., editors, *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Laicher, S., Kurtyigit, S., Schlechtweg, D., Kuber, J., and Schulte im Walde, S. (2021). Explaining and improving BERT performance on lexical semantic change detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 192–202, Online. Association for Computational Linguistics.

Lima-Lopes, R. E. d. (2025). BrPoliCorpus: Brazilian political corpus. Repositório de Dados de Pesquisa da Unicamp. DOI: 10.25824/REDU/YCFPIV. Available at: `https://redu.unicamp.br/citation?persistentId=doi:10.25824/redu/YCFPIV`. Accessed: May 15, 2026.

Lipani, A., Lupu, M., Hanbury, A., and Aizawa, A. (2015). Verboseness fission for BM25 document length normalization. In *Proceedings of the 2015 International Conference on the Theory of Information Retrieval*, pages 385–388, Northampton, Massachusetts, USA. Association for Computing Machinery.

Montariol, S., Martinc, M., and Pivovarova, L. (2021). Scalable and interpretable semantic change detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4642–4652, Online. Association for Computational Linguistics.

Osório, T. F. and Cardoso, H. L. (2025). Historical Portuguese corpora: a survey. *Language Resources and Evaluation*, 59:1797–1832.

Periti, F. and Tahmasebi, N. (2024). A systematic comparison of contextualized word embeddings for lexical semantic change. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4262–4282, Mexico City, Mexico. Association for Computational Linguistics.

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., and Tahmasebi, N. (2020). SemEval-2020 task 1: Unsupervised lexical semantic change detection. In Herbelot, A., Zhu, X., Palmer, A., Schneider, N., May, J., and Shutova, E., editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: Pretrained BERT models for Brazilian Portuguese. In *Intelligent Systems – BRACIS 2020*, pages 403–417, Rio Grande, Brazil. Springer.

Tahmasebi, N., Borin, L., Jatowt, A., Xu, Y., and Hengchen, S., editors (2021). *Computational Approaches to Semantic Change*. Language Science Press, Berlin.

Autonomous Fact-Checking: Integrating Local Inference and Adversarial NLP Attacks

Rômulo Henrique Nascimento Duarte[1], **Humberto Nunes de Lira**[1], **Vivianny Khatly Medeiros Pereira**[1], **Isaac Sebastian Lima de Araújo**[1], **Allan Gabriel da Cunha Vasconcelos**[1], **Yuri de Almeida Malheiros Barbosa**[1]

[1]Universidade Federal da Paraíba (UFPB)
João Pessoa – PB – Brazil

romulohenri000@gmail.com, nuneshumberto120@gmail.com, vkhatly@gmail.com
sisaac933@gmail.com, allan.vasconcelos@academico.ufpb.br, yuri@ci.ufpb.br

Abstract. *Digital misinformation threatens democratic discourse, especially in low-resource languages like Brazilian Portuguese. We present a fact-checking pipeline comparing structured claim decomposition (Claimify) against full-text extraction across four retrieval-augmented classifiers:* `gemini-2.5-flash-lite` *and three local open-weight models (Llama-3.1-8B, Qwen2.5-7B, Gemma-2-9B). Robustness is assessed under character-level and synonym perturbations. On a pilot sample, the API-based model shows a preliminary macro-F1 advantage for decomposition, which the three local models fail to replicate consistently. We also observe a systematic bias toward the false label across models.*

1. Introduction

Digital misinformation threatens democratic institutions and public discourse at an unprecedented scale, a challenge compounded in low-resource linguistic contexts such as Brazilian Portuguese by a scarcity of specialized tools and annotated resources [Guo et al. 2022, Silva et al. 2025]. Automated fact-checking has emerged as a promising direction to scale verification efforts beyond human-centric operations, typically framing the task as a classification problem in which a model assigns a veracity label to a claim, often supported by retrieved evidence [Thorne et al. 2018, Guo et al. 2022]; while English benefits from mature benchmarks [Wang 2017, Schlichtkrull et al. 2023], Brazilian Portuguese resources remain largely limited to binary, statement-only datasets [Silva et al. 2020, Garcia et al. 2022], with claim-level annotation efforts such as ClaimPT [Campos et al. 2026] only beginning to emerge for Portuguese. Beyond this data scarcity, two technical challenges limit existing systems' reliability: the quality of claim extraction directly affects downstream verification, since noisy or under-specified claims mislead classifiers regardless of their capacity [Metropolitansky et al. 2025, Guo et al. 2022]; and although adversarial robustness has received growing attention in NLP broadly [Morris et al. 2020, Jin et al. 2020], fact-checking systems are still seldom evaluated under adversarial conditions in low-resource languages, leaving their robustness to deliberate textual perturbations largely unknown in these settings. This paper investigates the following research question: **does structured claim decomposition improve the robustness and accuracy of LLM-based fact verification for Brazilian Portuguese, under both clean and adversarially perturbed inputs?** Our hypothesis is that decomposing documents into atomic, self-contained claims produces higher-quality classifier inputs than unstructured extraction. To investigate this, we build a deduplicated corpus from the *Aos Fatos* fact-checking portal, construct a balanced binary benchmark whose

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 113–125
October 19-22, 2026 ©2026 Association for Computational Linguistics

false class is generated through NLI-validated logical negation, and evaluate a retrieval-augmented classification pipeline under controlled surface-level perturbations. Our objectives are: (i) to compare structured claim decomposition (*Claimify*) against a full-text extraction baseline on an identical set of documents; (ii) to quantify the robustness of both configurations under character-level and synonym-substitution attacks at two intensity levels; and (iii) to characterize systematic error patterns of the resulting classifiers.

The main contributions of this work are: (i) a **controlled comparison of structured claim decomposition against full-text extraction**, evaluated on an identical set of documents under a shared retrieval-augmented setup, isolating the extraction strategy as the sole varying factor; (ii) a **cross-model evaluation spanning an API-based model and three local open-weight models**, showing that the preliminary decomposition advantage observed for the API-based model does not replicate consistently across smaller, locally deployed models, an open question we discuss in Section 4; and (iii) an **adversarial robustness protocol for Brazilian Portuguese fact-checking**, combining controlled surface-level perturbations with NLI-validated semantic preservation gates and a claim-level logical negation procedure used to construct a balanced verification benchmark.

2. Related Work

Automated fact-checking has been widely framed as a natural language inference or classification task, in which a model assigns a veracity label to a given claim [Thorne et al. 2018, Guo et al. 2022]. Early systems relied on structured knowledge bases; more recent approaches leverage pre-trained and, increasingly, large language models prompted in-context. LIAR [Wang 2017] established a six-class benchmark for statement-level verification, and AVERITEC [Schlichtkrull et al. 2023] extended the paradigm to evidence-grounded, multi-hop verification. For Brazilian Portuguese, existing work has largely been confined to binary classification on datasets derived from news portals and social media [Silva et al. 2020, Garcia et al. 2022]. A recent step towards richer Portuguese resources is ClaimPT [Campos et al. 2026], a dataset of European Portuguese news articles annotated for factual claims at the span level (1,308 articles and 6,875 annotations); it targets claim *detection*, whereas our work focuses on downstream *verification* and its robustness, and addresses the Brazilian rather than European variant, narrowing the gap between evidence-grounded verification benchmarks in English and what is available for low-resource settings [Zuccon et al. 2022]. Poorly specified claims propagate noise into downstream classifiers regardless of their capacity [Guo et al. 2022]; we integrate *Claimify* [Metropolitansky et al. 2025], an LLM-based method that decomposes documents into atomic, self-contained claims while explicitly handling referential and structural ambiguity, and compare it against a baseline extracting claims from the full document text without decomposition or disambiguation. Retrieval-Augmented Generation (RAG) addresses a fundamental limitation of purely parametric models: their inability to access up-to-date or domain-specific evidence [Lewis et al. 2020]. Early RAG-based fact-checking systems retrieved evidence from structured knowledge bases or Wikipedia [Thorne et al. 2018], while more recent approaches index domain-specific corpora [Pan et al. 2023]. All four classification arms in our pipeline are retrieval-augmented over the same locally hosted index, ensuring that differences reflect the claim extraction strategy rather than asymmetric access to evidence; our results, however, indicate that re-

trieval access alone does not make the extraction strategies behave uniformly across model families (Section 4). Foundational work exposed the fragility of NLP models to character-level perturbations and synonym substitutions [Morris et al. 2020, Jin et al. 2020], and a growing body of research now examines attacks tailored to automated fact-checking specifically [Liu et al. 2025]; while this line of work evaluates robustness mainly in English, we apply LLM-generated surface-level perturbations, at two controlled intensity levels and validated by NLI-based semantic gates, to Portuguese claims.

3. Methodology

This section describes the four main components of the proposed fact-checking pipeline: dataset construction, claim extraction, veracity classification, and adversarial robustness evaluation.

3.1. Dataset Construction

The primary corpus, hereafter DATASET-AOSFATOS, was built through a multi-stage web scraping pipeline targeting the *Aos Fatos* fact-checking agency, covering 20 thematic pages of the "public policies" (*"políticas públicas"*) channel, using rotating headers and a mandatory delay between requests to respect the portal's infrastructure. For each article, we prioritize extraction of embedded `ClaimReview` JSON-LD metadata over plain HTML parsing and isolate the main content container to extract the full body text, discarding articles lacking verifiable or programmatically mappable content. An integrity audit revealed 53 duplicate records (repeated `id_documento` with identical content) among 1,047 collected rows; since duplicates would be twice as likely to be selected under stratified sampling, biasing every downstream stage, all were removed prior to sampling, yielding a final pool of **994 unique documents** (crawler implementation details are in our repository). The final corpus was serialized into a `.jsonl` format with document IDs, source URLs, titles, and full report text, the input to all claim extraction configurations in Section 3.2.

Verification benchmark and label construction. Unlike our earlier formulation, veracity labels are not taken from the portal's `reviewRating` verdicts, which apply to the article as a whole, not individual claims. Instead, we construct a balanced binary benchmark at the claim level: canonical claims constitute the *true* class, and their *false* counterparts are generated through controlled logical negation, validated by bidirectional NLI contradiction gates (Section 3.4), yielding one *false* instance per *true* claim and a 50/50 balance verified programmatically in every cell.

Sampling. From the pool of 994 unique documents, we sample **50 documents** using a fixed seed (42) with stratification by document length; documents used as few-shot demonstrations are kept outside the analysis set.

3.2. Claim Extraction

Raw fact-checking reports contain multiple sentences, background context, and narrative elements. Transforming this text into well-formed, verifiable claims is the first stage of

the pipeline, and the central experimental variable of this work. In both configurations described below, claims are extracted from the **full body text** of each report; headlines are never used as claim proxies. This holds for every stage of the pipeline, including classification (Section 3.3).

We evaluate two extraction conditions. **Baseline extraction**: an LLM is prompted to extract the factual claims present in the full document text in a single pass, without any decomposition or disambiguation stages, a straightforward, low-engineering approach. **Claimify**: we apply *Claimify* [Metropolitansky et al. 2025], an LLM-based method that decomposes the document into atomic, self-contained claims through dedicated selection, disambiguation, and decomposition stages, explicitly handling referential and structural ambiguity; following the method's staged design, we execute these stages in batched form per document (three LLM calls per document per condition), a documented adaptation of the original per-sentence formulation.

Both conditions use the same underlying LLM (Section 3.3) and receive identical inputs, so observed differences are attributable to the extraction strategy alone. The two strategies produce structurally different outputs: the baseline extracts on average 13.2 claims per document (660 total), while Claimify's decomposition yields 52.4 (2,621 total), reflecting its finer-grained segmentation. Extracted claims, hereafter *canonical claims*, constitute the *true* class and the source material for the perturbation and negation procedures (Section 3.4); document-level comparability between arms is enforced by the intersection protocol defined there.

3.3. Veracity Classification

We evaluate veracity classification through In-Context Learning (ICL) strategies over a retrieval-augmented pipeline. Every classification arm, regardless of extraction strategy, receives retrieval-augmented context from the same locally hosted index, ensuring symmetric access to evidence across all experimental conditions.

3.3.1. Classification Model

All pipeline stages requiring an LLM (claim extraction, perturbation, negation, and classification) use `gemini-2.5-flash-lite` at temperature 0, accessed through its floating alias since no dated snapshot was available at experiment time; we document this as an explicit reproducibility deviation, anchoring reproducibility to each run's execution date rather than an immutable identifier (prompts, hyperparameters, and execution metadata are available in our repository). The classifier receives the claim, retrieved evidence passages (Section 3.3.3), and a directive specifying the two allowed labels and output format, producing a brief reasoning followed by a final verdict line from which the label is parsed; as established in Section 3.2, claims under classification originate from the full body text of the source reports in both arms.

3.3.2. Local Models

To characterize how extraction strategy interacts with smaller, open-weight models, we evaluate three locally deployed instruction-tuned models, **Llama-3.1-8B-Instruct**, **Qwen2.5-7B-Instruct**, and **Gemma-2-9B-it**, served via the Hugging Face `transformers` pipeline with 4-bit quantization (BitsAndBytes) on a single GPU, ge-

nerating up to 250 new tokens per response at temperature 0. Local models reuse the same retrieval context (Section 3.3.3), prompt contracts, and parsing policy (Section 3.3.4) as the API-based model, with each surviving claim yielding one classification call per surface condition plus one for its negated counterpart (Section 3.4.4). Due to the cost of local generation, they are evaluated on half the pilot's surviving claims (44 of 89 baseline; 227 of 455 Claimify, same seed), yielding 3,252 raw calls per model; since this subsample is smaller than, and independently drawn from, the full pilot used for the API-based model, we report local results separately rather than in a direct comparison table (Section 7).

3.3.3. Retrieval-Augmented Generation

Fact-checking reports from DATASET-AOSFATOS are segmented into overlapping passages using a sliding window of 3 sentences with step 1 (via NLTK's Portuguese Punkt tokenizer) and indexed in a locally hosted, persistent ChromaDB instance; the index thus contains the same *Aos Fatos* articles from which the canonical claims are extracted, a design choice we control for explicitly. For each claim, passages originating from its own source document are excluded from retrieval at query time, preventing the classifier from trivially retrieving the report that generated the claim, an exclusion verified by automated audit for every run. Passages and queries are encoded with `intfloat/multilingual-e5-large`, pinned to a specific revision, with normalized embeddings; the retriever fetches a buffered candidate set, discards source-document passages, deduplicates near-identical windows from the same document (windows sharing two or more sentence indices), and returns the top-5 remaining passages, injected into the prompt as supporting context.

3.3.4. Prompting and Output Parsing

All configurations are evaluated under two prompting regimes: **zero-shot**, in which the model receives the claim, retrieved context, and a task description with no labeled examples; and **few-shot**, in which four labeled demonstrations (claim, short reasoning, final label) are prepended, drawn from documents outside the analysis set to prevent contamination, a constraint verified programmatically.

The classifier outputs one of the two target labels; any unparsable response, including explicit declines to commit citing insufficient evidence, is counted as a classification *error* under a policy fixed before the experiments, reflecting the deployment view that a pipeline failing to produce a verdict has failed at its task. We report this failure rate (`parse_fail`) separately in all result tables.

3.4. Adversarial Robustness Evaluation

To assess the pipeline's sensitivity to deliberate textual perturbations, we apply surface-level adversarial attacks at two intensity levels and evaluate classification performance across all resulting conditions. Unlike our earlier formulation, logical negation is not treated as an attack: it is the mechanism that generates the *false* class of the benchmark, described below. All perturbations and negations are generated by the same LLM as the rest of the pipeline (Section 3.3.1), at temperature 0.

3.4.1. Surface-Level Attack Strategies

Perturbations are applied to the full document body before claim extraction, so that each extraction arm processes clean and attacked versions of the same documents. Two strategies are used, each at two intensity levels: **character-level perturbations (typos)**, realistic typographic errors (swaps, insertions, deletions) controlled by an edit budget of one per 16 content words (*light*) or one per 8 (*strong*); and **synonym substitution**, replacing content words with semantically similar alternatives at a rate of 15% (*light*) or 30% (*strong*). Together with the unperturbed *identity* condition, which serves as the control, this yields five surface conditions per document per arm.

3.4.2. False-Class Generation via Logical Negation

The *false* class is generated by logical negation applied to each canonical claim *individually*: each claim is rewritten so that its factual core is inverted while remaining fluent, and the resulting claim receives the label *false* by construction (labels under negation are *inverted*, never preserved). Because negation operates claim-by-claim, each original claim maps one-to-one to its negated counterpart, requiring no post-hoc matching. Every negated claim must pass a bidirectional contradiction gate, a multilingual NLI model (`MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7`) classifying the original claim as contradicting the negated version *and* vice versa; claims failing this gate are excluded, ensuring every *false* instance is a genuine semantic inversion rather than a superficial rewording. This per-claim procedure achieves gate approval rates of 90.0% (baseline) and 91.8% (Claimify), well above the 12–19% approval of an earlier document-level attempt, where the model frequently left individual assertions unchanged, motivating the per-claim formulation.

3.4.3. Alignment, Validation Gates, and Intersection Protocol

For the surface conditions, claims extracted from attacked documents must be matched to their clean counterparts before any comparison is meaningful. We align each attacked-document claim to canonical claims via embedding cosine similarity (same `multilingual-e5-large` model as the retrieval index), with a threshold of $\tau = 0.90$ fixed by inspection of the pilot score distribution (± 0.05 sensitivity showed near-constant survival rates, so the threshold is not critical); aligned pairs must also pass an NLI-based **semantic preservation gate**, verifying the perturbation did not alter the claim's meaning beyond the intended surface-level noise. The final evaluation set is the **intersection** of claims surviving alignment and gating across all conditions, guaranteeing every configuration (extraction arm $\times$ prompting regime $\times$ surface condition) is evaluated on the same content, eliminating survivorship differences as a confound. We denote by σ the fraction of sampled documents in this intersection ($\sigma = 0.58$, 29 of 50 in the pilot); within each arm, every cell is exactly balanced between *true* and *false* instances (50/50), verified programmatically.

3.4.4. Evaluation Protocol and Metrics

For each evaluation cell we report accuracy, **macro F1-score**, per-class F1 (*true*/*false*), and `parse_fail` (Section 3.3.4). Since every cell is balanced 50/50 by construction, macro and weighted F1 coincide, so we report macro F1 uniformly as the primary metric.

Robustness to a given attack is the degradation relative to identity, $\Delta\text{F1} = \text{F1}_{\text{id}} - \text{F1}_{\text{attacked}}$, where smaller ΔF1 indicates greater robustness.

4. Results and Discussion

All results below come from the pilot run: $D = 50$ documents sampled as described in Section 3.1. Given the sample size, we present these findings as preliminary evidence rather than definitive conclusions, and we dimension every claim accordingly.

4.1. Extraction and Survival Statistics

As established in Section 3.2, the two arms yield 660 and 2,621 canonical claims, respectively. After alignment, validation gates, and the intersection protocol (Section 3.4.3), 89 baseline and 455 Claimify claims survive, drawn from the 29 documents in the global intersection ($\sigma = 0.58$); Claimify's finer-grained decomposition thus yields both more claims per document and a higher survival volume, though survival *rates* per condition vary by attack type, as shown next.

4.2. Alignment and Validation Gates

Table 1. Alignment and gate-pass rates, baseline vs. Claimify (pilot, $D = 50$; $n = 660$/2,621 claims per condition). Aligned/gate pass/cos as defined in Section 3.4.3.

	Baseline			Claimify		
Condition	Aligned	Gate pass	Cos	Aligned	Gate pass	Cos
negation	1.000	0.900	0.987	1.000	0.918	0.978
syn strong	0.874	0.361	0.983	0.890	0.481	0.980
syn light	0.898	0.388	0.983	0.822	0.436	0.975
typo strong	0.855	0.533	0.990	0.863	0.590	0.987
typo light	0.830	0.477	0.989	0.886	0.589	0.990

Table 1 reports, per arm and condition, the fraction of claims aligned at $\tau = 0.90$ (Aligned), passing the semantic gates (Gate pass), and mean cosine similarity of aligned pairs (Cos). Negation passes at high rates in both arms (0.900 and 0.918), confirming the reliability of the per-claim procedure (Section 3.4.2). Surface conditions pass at substantially lower rates (0.36–0.59), as expected: the gates are deliberately conservative, discarding perturbed claims whose meaning drifted beyond surface noise. Typo conditions survive at higher rates than synonym conditions in both arms, consistent with synonym substitution being the more invasive perturbation.

4.3. Classification Results

Table 3 reports classification results per arm and prompting regime, averaged over the five surface conditions; the full per-cell results for all twenty evaluation cells are available in our repository. Three patterns emerge. **Claimify shows a preliminary advantage.** Averaged over surface conditions, the Claimify arm outperforms the baseline by 2.5 macro-F1 points under few-shot (0.667 vs. 0.642) and 3.0 points under zero-shot (0.624 vs. 0.594). Given the modest number of surviving baseline claims (89 claims, 178 instances per cell),

we interpret this as preliminary evidence favoring structured decomposition, in line with the open debate on whether claim decomposition helps or hinders downstream verification [Hu et al. 2025], not as a conclusive result.

Few-shot consistently outperforms zero-shot in both arms and in every surface condition, with gains of 4.3–4.8 macro-F1 points. This contrasts with our earlier formulation, in which few-shot prompting sometimes degraded performance on noisy inputs, and suggests that with retrieval-augmented context and controlled claim quality, in-context demonstrations provide a stable benefit.

Robustness to surface attacks is high in both arms. Across all twenty cells, degradation relative to identity is small throughout ($|\Delta\text{F1}| \leq 0.028$), indicating the surviving, gate-validated perturbations rarely flip classifier decisions. We emphasize the selection effect: these are precisely the perturbations that *preserved* claim semantics, so this measures robustness to meaning-preserving noise, not arbitrary corruption.

Table 2. Per-cell classification results (pilot, $D = 50$), by arm, surface condition, and prompting regime (fs/zs). id = identity (unperturbed control). Each cell is balanced 50/50, with $n = 178$ instances per cell in the baseline arm and $n = 910$ in the Claimify arm. F1t/F1f = F1 for *true*/*false*; pf = parse_fail.

		Macro F1		F1t		F1f		pf	
Arm	**Surface**	fs	zs	fs	zs	fs	zs	fs	zs
baseline	id	0.651	0.596	0.582	0.484	0.721	0.708	0.039	0.011
baseline	syn strong	0.641	0.596	0.571	0.484	0.711	0.708	0.028	0.011
baseline	syn light	0.668	0.602	0.611	0.496	0.725	0.708	0.028	0.006
baseline	typo strong	0.615	0.587	0.529	0.472	0.701	0.702	0.034	0.006
baseline	typo light	0.636	0.587	0.561	0.472	0.711	0.702	0.034	0.006
claimify	id	0.667	0.625	0.605	0.545	0.730	0.706	0.027	0.042
claimify	syn strong	0.662	0.624	0.598	0.540	0.726	0.708	0.026	0.047
claimify	syn light	0.671	0.626	0.613	0.545	0.729	0.707	0.022	0.043
claimify	typo strong	0.669	0.619	0.607	0.536	0.731	0.702	0.029	0.040
claimify	typo light	0.667	0.625	0.605	0.545	0.729	0.706	0.026	0.042

Table 3. Average classification results per arm and prompting regime (pilot, $D = 50$), averaged over the five surface conditions.

Arm	Prompt	Acc	Macro F1	F1_{true}	F1_{false}	parse_fail
baseline	fs	0.646	0.642	0.571	0.714	0.033
baseline	zs	0.623	0.594	0.482	0.706	0.008
claimify	fs	0.669	0.667	0.606	0.729	0.026
claimify	zs	0.628	0.624	0.542	0.706	0.043

4.4. Systematic Error Patterns

Two error patterns hold across both arms and all conditions.

Systematic bias toward *false*. F1 on the *false* class (0.70–0.73) is consistently higher than on *true* (0.47–0.61) in every cell; since cells are balanced by construction, this reflects classifier behavior, not class imbalance. A plausible, untested factor is the retrieved context itself: fact-checking reports frequently contain explicit debunking language, which may prime the classifier toward *false* verdicts regardless of the claim.

Non-committal responses concentrate in the Claimify arm. The `parse_fail` rate is consistently higher for Claimify under zero-shot (4.0–4.7%) than baseline (0.6–1.1%); inspection of raw responses confirms these are explicit declines to verdict, not parsing artifacts, more frequent for the atomic, decontextualized claims Claimify produces, whose evidentiary requirements are harder to satisfy within the top-5 passages. Under our pre-registered policy (Section 3.3.4) these count as errors, so the Claimify advantage above is achieved *despite* this penalty, a tension with decomposition granularity aligning with [Hu et al. 2025] and, in our view, the most consequential open question for pipelines of this kind.

4.5. Local Models: Cross-Model Comparison

Table 4. Local models: results by arm, prompting regime, and metric, averaged over five surface conditions (44 baseline / 227 Claimify claims; 88/454 instances per cell). F1t/F1f = F1 for *true*/ *false*; pf = parse_fail.

		Macro F1		F1t		F1f		pf	
Model	**Arm**	fs	zs	fs	zs	fs	zs	fs	zs
Gemma-2-9B	baseline	0.636	0.655	0.545	0.553	0.727	0.757	0.186	0.123
Gemma-2-9B	Claimify	0.560	0.587	0.546	0.555	0.573	0.620	0.340	0.277
Llama-3.1-8B	baseline	0.422	0.531	0.296	0.400	0.549	0.662	0.434	0.177
Llama-3.1-8B	Claimify	0.430	0.539	0.321	0.430	0.538	0.648	0.467	0.152
Qwen2.5-7B	baseline	0.586	0.504	0.489	0.314	0.684	0.694	0.080	0.005
Qwen2.5-7B	Claimify	0.546	0.516	0.440	0.338	0.651	0.694	0.075	0.007

Table 4 reports classification results for the three local models, averaged over the five surface conditions, on the reduced subsample described in Section 3.3.2 (44 baseline / 227 Claimify claims; 88/454 balanced instances per cell). Three patterns emerge, two extending the API-based model's findings and one specific to local deployment.

No consistent effect of structured decomposition emerges. Gemma-2-9B favors the baseline in both regimes (-7.6 and -6.8 points), Llama-3.1-8B shows negligible differences ($\leq$0.8 points), and Qwen2.5-7B favors baseline under few-shot but reverses under zero-shot, unlike the API-based model's consistent advantage (Section 4.3).

Non-committal responses vary far more across local models than for the API-based model. `parse_fail` ranges from under 1% (Qwen2.5-7B, zero-shot) to 46.7% (Llama-3.1-8B, Claimify, few-shot), an order of magnitude wider than `gemini-2.5-flash-lite`'s 2.7–5.7%. Llama-3.1-8B and Gemma-2-9B are especially affected under few-shot (43–47% and 19–34%), while Qwen2.5-7B stays close to the API-based model's range. Since local generation is capped at 250 tokens (Section 3.3.2), we cannot fully disentangle genuine uncertainty from response truncation, a confound discussed further in Section 7.

The bias toward *false* is present in all three local models, more pronounced in two. Llama-3.1-8B and Qwen2.5-7B show large true/false F1 gaps under zero-shot (Llama: 0.40 vs. 0.66; Qwen: 0.31–0.34 vs. 0.69), wider than for the API-based model, while Gemma-2-9B shows a smaller gap under baseline but comparable under Claimify. This asymmetry, robust across all four models, is discussed in Section 4.4.

5. Conclusion

This paper presented a fact-checking pipeline for Brazilian Portuguese that addresses two underexplored questions: whether structured claim decomposition improves classification quality, and how robust the resulting classifiers are, both to adversarial perturbation and across model families. By constructing a deduplicated corpus from *Aos Fatos*, building a balanced verification benchmark through NLI-validated logical negation, and evaluating four retrieval-augmented models under a shared protocol, we obtained a more nuanced picture than either extreme would suggest. For the API-based model, structured decomposition via *Claimify* shows a consistent, if preliminary, macro-F1 advantage over full-text extraction (Section 4.3), an advantage that does **not** replicate consistently across the three local, open-weight models evaluated under the same protocol (Section 4.5). Rather than supporting a uniform claim about decomposition, these results suggest that its benefit is conditional on model capacity or training, an open question we did not anticipate at the outset and that we believe merits targeted follow-up work. Two further patterns emerged across all four models: a systematic bias toward the *false* label, more pronounced in the local models, and non-committal responses (`parse_fail`) that varied far more widely among local models than for the API-based model, a spread plausibly confounded by our fixed local generation budget rather than genuine differences in model uncertainty. Taken together, this work argues against treating claim decomposition, or open-weight local deployment, as unconditionally beneficial for retrieval-augmented fact verification; their effects appear to interact with model-specific factors a single pilot cannot disentangle, an invitation for larger-scale, model-diverse evaluations rather than a negative result.

6. Future Work

The most immediate next step is a **controlled, equal-scale cross-model comparison** of all four models on the same survivor set and token budget, disentangling whether local models' failure to replicate the Claimify advantage is genuine or an artifact of the reduced subsample and generation budget used here (Section 7). We also plan to apply bootstrap confidence intervals to the pilot results, quantifying how much of the 2.5–3.0 point Claimify advantage is attributable to sampling variability; isolate the contribution of retrieval (RAG on/off, varying k) from that of decomposition; and investigate the systematic bias toward *false*, for instance by filtering verdict-bearing sentences from retrieved context. Further directions include expanding DATASET-AOSFATOS to other agencies, a human-reliability audit of the canonical claim set, and evaluating additional open-weight model families.

Reproducibility Statement

All data, code, and configuration needed to reproduce this work, corpus construction scripts, full prompt texts, the configuration file (seed, thresholds, retrieval parameters), and evaluation code, are publicly available in our repository, along with basic run instructions. As documented in Section 3.3.1, the classification LLM is the only component not pinnable to an immutable identifier, so its execution date is recorded instead. All artifacts are archived at `https://doi.org/10.17605/OSF.IO/YZ736`.

7. Limitations

Sample size and local model comparability. Results come from a pilot of 50 documents, with the baseline arm retaining only 89 surviving claims. The reported Claimify advantage for the API-based model (2.5–3.0 macro-F1 points) is within a range where sampling variability cannot be ruled out; we frame it as preliminary throughout. The three local models are evaluated on an independently sampled, smaller subset (half the pilot's survivors), so their results cannot be directly compared to `gemini-2.5-flash-lite`'s in a single table. Equal-scale, cross-model evaluation with confidence intervals is the immediate next step.

Confounded parse failure rates for local models. Local generation was capped at 250 tokens for GPU memory and runtime reasons. `parse_fail` rates range from under 1% to over 46%, far wider than the API-based model's range, and markedly higher under few-shot for two of the three models. We cannot distinguish genuine model uncertainty from response truncation before the verdict line, limiting how much of the local models' failure to replicate the Claimify effect owes to extraction strategy versus this generation budget. **Scope.** All four models share the same retrieval setup ($k = 5$) and embedding model; we did not run ablations isolating the contribution of retrieval or architecture. The API-based model is accessed through a floating alias, with reproducibility anchored by execution date rather than an immutable identifier. DATASET-AOSFATOS is scraped from a single agency and bounded to its public-policy channel, which risks institutional and topical bias; canonical claims are extracted from reports that may quote the very misinformation they debunk, a failure mode our gates do not screen for and that we did not audit with human review.

Ethical Considerations

The dataset was constructed by scraping publicly available content from the *Aos Fatos* fact-checking portal, with request delays to minimize server impact and no collection of personally identifiable information. Automated fact-checking systems, including the pipeline proposed here, should not be used as sole arbiters of truth: classification errors are inevitable, and human oversight remains essential for consequential decisions. The adversarial perturbation and negation procedures generate synthetic false claims strictly for controlled evaluation, disclosed with a defensive intent to expose pipeline vulnerabilities; we acknowledge such techniques could nonetheless be misused to evade fact-checking or fabricate misleading content at scale. Finally, the language models used may reflect biases in their pre-training data, and outputs should be interpreted with appropriate caution in politically sensitive contexts.

Use of Artificial Intelligence

This work used generative AI tools as writing and engineering assistants: **Claude Fable 5** and **Claude Sonnet 4.6** (Anthropic) supported drafting and revising manuscript sections and the implementation of the experimental pipeline, under the authors' direction. All design decisions, data, analyses, and conclusions remain the authors' sole responsibility; they reviewed and validated all AI-assisted content.

Referências

Campos, R., Sequeira, R., Nerea, S., Cantante, I., Folques, D., Cunha, L. F., Canavilhas, J., Branco, A., Jorge, A., Nunes, S., Guimarães, N., and Silvano, P. (2026). ClaimPT: A portuguese dataset of annotated claims in news articles. In *Advances in Information Retrieval (ECIR 2026)*, volume 16486 of *Lecture Notes in Computer Science*, pages 544–560. Springer.

Garcia, D., Silva, A., and Santana, C. (2022). Fake news detection in brazilian portuguese: Datasets and current challenges. In *Proceedings of the Brazilian Symposium on Information Systems*, pages 115–122.

Guo, Z., Schlichtkrull, M., and Vlachos, A. (2022). From ready-to-wear to ready-to-match: A survey on automated fact-checking. *ACM Computing Surveys*, 55(6):1–34.

Hu, Q., Long, Q., and Wang, W. (2025). Decomposition dilemmas: Does claim decomposition boost or burden fact-checking performance? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6313–6336.

Jin, D., Jin, Z., Zhou, J. T., and Szolovits, P. (2020). Is BERT really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8018–8025.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Lewis, V., Riedel, S., and Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, pages 9459–9474.

Liu, F., Abuadbba, A., Moore, K., Nepal, S., Paris, C., Wu, J., Yang, J., and Sheng, Q. Z. (2025). Adversarial attacks against automated fact-checking: A survey. *arXiv preprint arXiv:2509.08463*.

Metropolitansky, I., Goldstein, A., and Shahaf, D. (2025). Claimify: Decomposing documents into atomic verifiable claims using large language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 412–426.

Morris, J., Lifland, E., Yoo, J. Y., Grigsby, J., Jin, D., and Qi, Y. (2020). TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126.

Pan, J., Wang, X., and Gao, Y. (2023). A survey on retrieval-augmented text generation for large language models. *arXiv preprint arXiv:2304.09400*.

Schlichtkrull, M., Guo, Z., and Vlachos, A. (2023). AVeriTeC: A dataset for real-world fact checking with evidence. *arXiv preprint arXiv:2305.13117*.

Silva, J., Santos, R., Almeida, T., and Pardo, T. (2020). Fake.Br corpus: A corpus for fake news detection in brazilian portuguese. In *Proceedings of the International Conference on Computational Processing of Portuguese*, pages 416–425.

Silva, R. M., Amamou, H., Ferraz, L. B. S., da Silva, F. K. A., and Avila, A. R. (2025). Fake news detection in portuguese under large language model-generated content. *Journal of the Brazilian Computer Society*, 31(1):1149–1166.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. (2018). FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the Human Language Technologies: NAACL*, pages 809–824.

Wang, W. Y. (2017). Liar, liar pants on fire: A new benchmark dataset for fake news detection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 422–426.

Zuccon, G., Koopman, B., and Palotti, J. (2022). The landscape of fact-checking resources for low-resource languages. *Journal of Artificial Intelligence Research*, 74:1435–1461.

A Comparison of Commonly Used Automatic Evaluation Metrics for Open-Ended Tasks in Portuguese

Eduardo D. Faé, Dennis Giovani Balreira, Viviane Pereira Moreira

Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre – RS – Brazil

{edfae,dgbalreira,viviane}@inf.ufrgs.br

Abstract. *Open-ended tasks in Natural Language processing are characterized by the absence of a fixed set of outputs or short, predetermined responses. Typical examples include open-domain question answering and summarization. The automatic evaluation of these tasks is challenging, and existing metrics suffer from important limitations. Thus, many works have already been proposed to evaluate the performance of such metrics. However, the great majority of these works were conducted in English, leaving a gap for analyses in other languages, such as Portuguese. This work investigates the performance of some of the most popular automated intrinsic evaluation metrics for open-ended tasks by analyzing their correlation with human judgment. Using both the Summarization and Question Answering tasks, this study compares traditional n-gram-based metrics with metrics based on contextual embeddings generated by Pre-trained Language Models (PLMs), performing all analyses using models and datasets available for Portuguese. Results indicate that while PLM-based metrics generally show slightly higher correlation with human perception, their performance is highly reliant on the quality of the models used. Still, none of the evaluated metrics displayed a strong correlation with human judgments, suggesting the need for more robust metrics.*

1. Introduction

Open-ended tasks are problems designed so that they may contain more than one correct solution or be solved in more than one unique way. These tasks typically allow for a diversity of valid outputs and interpretations and are commonly present in areas that require creativity and critical thinking. They are extremely common in Natural Language Processing (NLP), as natural languages are inherently dynamic, volatile, diverse, and context-dependent.

These tasks can be found everywhere, from translating a sentence to writing an article. In NLP, two common open-ended tasks are summarization and question answering (QA). Both tasks recently reached new levels of importance, with the diffusion of Large Language Models (LLMs) and popularization of newer Information Retrieval (IR) techniques, such as RAG [Lewis et al. 2020]. However, like many other open-ended tasks, they still share common problems in terms of evaluation.

In general, open-ended tasks are significantly more complex to verify and evaluate than closed-ended ones. This is a result of their unique ability to produce multiple semantically correct outputs for a single input, which poses challenges when analyzing model performance. Many metrics have been proposed to address this issue, some of which have

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 126–138
October 19-22, 2026 ©2026 Association for Computational Linguistics

become widely used as comparison tools in the NLP field. Traditional metrics generally follow the concept of n-grams [Shannon 1948] as a means of evaluating an output given a ground-truth, while newer metrics tend to rely on Pre-trained Language Models (PLMs) to try capturing the semantic similarity between a generated text and an expected answer.

For such, many works have sought to compare the performance of different metrics and examine their correlation with human evaluation [Liu et al. 2016, Deutsch et al. 2022, Liu et al. 2023]. With the translation task in particular having a great start, given the promotion of multiple large-scale evaluation campaigns aimed to measure the performance of early-adopted metrics, such as [NIST 2006, Koehn and Monz 2006]. However, when it comes to summarization, and especially QA, the number of works becomes a lot smaller [Fabbri et al. 2021, Junqueira and Moreira 2026]. Furthermore, most existing studies focus primarily on English datasets and models, leaving other languages, such as Portuguese, comparatively underexplored.

This work aims to perform these correlation analyses with models and datasets available for Portuguese. Still, this work focuses on two distinct types of open-ended tasks, summarization and QA, the latter being a task for which the literature remains comparatively limited. Additionally, we integrate real human evaluators to avoid introducing bias via tools like LLM-as-a-judge.

To achieve this, we collected evaluations from real human judges. With each of the two tasks having 40 generated answers, evaluated by four distinct judges, in four quality dimensions. This accounted for 160 evaluations per quality dimension for each task, totaling 1,280 human evaluations. The results showed that PLM-based metrics present slightly higher correlations with human judgments than n-gram-based ones. However, they appear to be extremely reliant on the quality of the PLM used. Still, both types of metrics failed to maintain high degrees of correlation with human judgments.

2. Background

Abstractive Approach. When it comes to summarization and QA, approaches generally fall into two categories: extractive and abstractive. While extractive approaches select and reproduce portions of the original text, abstractive approaches generate new text that may paraphrase, reorganize, or synthesize information from the source. In general, the abstractive approach is the more open-ended of the two, as it allows for a more diverse set of answers. Thus, will be the preferred approach for this work.

Intrinsic Evaluation. Evaluation methods in NLP can generally be divided into intrinsic and extrinsic approaches. Intrinsic evaluation assesses a model directly on its intended task using task-specific metrics, whereas extrinsic evaluation assesses the model by measuring its impact on the performance of a downstream application or end task. Performing intrinsic evaluation in open-ended tasks is particularly challenging, given the nature of such tasks, in which an output can be partially correct, and many distinct outputs can be correct, allowing for variations in writing styles, lexical choices, semantic paraphrases, textual structures, or levels of detail. Which lead to the proposal of a plethora of automated metrics, such as ROUGE [Lin 2004] and BERTScore [Zhang et al. 2019].

N-gram-based Metrics. n-gram-based metrics evaluate generated texts by measuring the overlap of contiguous word sequences between candidate and reference outputs. These

methods assume that higher lexical similarity indicates better generation quality, making them computationally efficient and easy to reproduce. However, in open-ended tasks, semantically correct answers may be expressed using different vocabulary or sentence structures, which limits the ability of n-gram-based approaches to fully capture meaning and contextual adequacy. Despite these limitations, such metrics remain widely adopted for evaluating NLP systems, with BLEU [Papineni et al. 2002], ROUGE [Lin 2004], and METEOR [Banerjee and Lavie 2005] among the most commonly used approaches.

PLM-based Metrics. PLM-based metrics evaluate generated texts by leveraging contextualized semantic representations obtained from language models. These metrics can capture semantic similarity even when the candidate and reference texts use different lexical choices or syntactic structures. As a result, PLM-based metrics generally achieve stronger correlations with human judgments in open-ended tasks. Nevertheless, they incur higher computational costs and still struggle to assess factual consistency or reasoning quality. Among the most prominent PLM-based evaluation metrics are MoverScore [Zhao et al. 2019], BERTScore [Zhang et al. 2019], and BARTScore [Yuan et al. 2021].

3. Related Work

Several studies have investigated the correlation between automatic evaluation metrics and human judgments for open-ended tasks. Back in 2005, the creators of METEOR [Banerjee and Lavie 2005] demonstrated that their technique achieved higher correlation with human judgments than BLEU [Papineni et al. 2002]. In the following years, large-scale evaluation campaigns, such as those conducted by NIST [NIST 2006] and WMT [Koehn and Monz 2006], began systematically comparing the most prominent metrics in machine translation against collected human annotations. This led to the popularization of these metrics, promoting their adoption for other open-ended NLP tasks as better automated evaluation methods.

Around 2015, analyses of these metrics in other open-ended tasks, such as summarization and dialogue generation, began to emerge. However, studies from this time, such as [Liu et al. 2016], still demonstrated limited correlation between traditional metrics and human evaluation, reporting extremely low results. Recent studies still show medium to low correlations between automated metrics and human judgments [Deutsch et al. 2021, Fabbri et al. 2021, Deutsch et al. 2022, Liu et al. 2023, Junqueira and Moreira 2026].

For the Summarization field, one of the most important studies was SummEval [Fabbri et al. 2021], which introduced a benchmark designed to investigate the correlation between automatic summary evaluation metrics and human judgment. In their analysis, they divided the evaluation of each summary into four quality dimensions: coherence, fluency, relevance, and consistency, aiming to capture different aspects that affect the quality of a summary. The authors tested a wide variety of metrics and pointed out that metrics based on n-grams, such as BLEU and ROUGE [Lin 2004], show low correlation with human perception, whereas PLM-based metrics present better results but still fall short of fully capturing the nuances and aspects present in human evaluation.

In contrast, comparatively fewer studies have investigated automatic evaluation for QA. Notable examples include: [Krishna et al. 2021], who demonstrated that the ROUGE-L metric provides limited insight for this type of task, as it can be artificially optimized without reflecting true answer quality; [Farea et al. 2025], which presents a broader

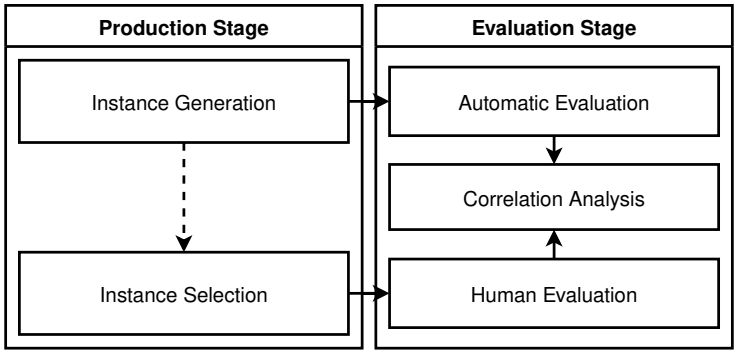
analysis of existing QA systems; and [Junqueira and Moreira 2026], which demonstrated how lexical metrics are limited in evaluating QA for Portuguese. The latter of which is among the few studies that analyzed the correlation between automatic metrics and human judgments in QA for non-English languages.

This work aims to expand the literature by addressing two main limitations: the lack of Portuguese-focused studies on the reliability of automatic intrinsic evaluation metrics for open-ended tasks and the lack of QA-focused analyses of metric correlations. Still, unlike recent studies such as [Junqueira and Moreira 2026], we focus on introducing a unified multi-task methodology spanning both summarization and QA, which enables a better analysis of the current capacities of automatic intrinsic evaluation metrics for open-ended tasks in Portuguese.

4. Methodology

This work is divided into five stages to better organize and present the methodology and its execution. Further explanations about each stage are provided throughout this section. Figure 1 provides an overview of this methodology. During the production stage, solutions are generated and selected; in the evaluation stage, solutions are evaluated by automated metrics and human judges, enabling a correlation analysis.

Figure 1. Overview of the methodology.



4.1. Instance Generation

To facilitate test replication and avoid bias in the use of an authorial dataset, both models and datasets were sourced from public repositories. Described below are the models and datasets chosen for the purposes of this work.

Models. When analyzing the correlation between a metric and human evaluation, it is essential to have a wide range of good and bad outputs. For this, we opted for smaller models, as they tend to be more unstable, better encompassing the desired quality range.

For Summarization, the model used was the Portuguese T5 for Abstractive Summarization [Paiola et al. 2022], with the fine-tuned version on the XL-Sum dataset [Hasan et al. 2021] being chosen. It was selected after a round of testing for being a small model that originated in an article published at a renowned conference.

In our search for small Portuguese models for the QA task, we failed to find open models that achieved a minimum standard and came from reliable sources. Thus,

for QA, the model selected was the smallest instruction-tuned version of Gemma 3 [Team et al. 2025]. This fulfills our goal of being a small model and, while not being primarily developed with Portuguese in mind, still has native support for Portuguese.

Datasets. The predominant language in NLP datasets is English; therefore, finding well-cataloged and curated datasets for languages other than English is a difficult task. It was no different for this work; with most of the tested datasets presenting problems, whether due to errors stemming from direct translation or poorly designed references. After several tests, the datasets selected are described below.

For Summarization, the dataset used was the CNN Dailymail Azure Pt[1]. This is a translated version of the CNN/DailyMail dataset [Nallapati et al. 2016], which provides a catalog of news articles from the media outlet, accompanied by written summaries used as references for calculating metrics. This dataset was chosen because it is a translation of one of the datasets used by SummEval [Fabbri et al. 2021], enabling a better comparison of the results in this work with those reported by SummEval.

For QA, the dataset used was FairytaleQA Translated Pt [Leite et al. 2024]. This dataset was created from the automatic translation of the open-source FairytaleQA dataset [Xu et al. 2022], which contains a compilation of short fairy tales and questions related to each tale, accompanied by their expected answers. The dataset was chosen for its ludic tone and focus on the abstractive approach, which led to a decrease in directly extracted answers when compared to most others, which tend to follow the SQuAD [Rajpurkar et al. 2016] pattern and focus on the extractive approach.

4.2. Automatic Evaluation

As a means of automatic evaluation, three of the most influential metrics from two of the most well-established approaches used for automatic intrinsic evaluation were selected. Table 1 presents each metric, the approach they are based on, and their year of creation.

Table 1. Description of the metrics evaluated by this work.

Metric	*Based on*	*Created in*
BLEU	n-grams	2002
ROUGE	n-grams	2004
METEOR	n-grams	2005
BERTScore	PLMs	2019
MoverScore	PLMs	2019
BARTScore	PLMs	2021

PLM-based metrics rely on strong Portuguese models to achieve good results. For this, the large version of BERTimbau [Souza et al. 2020] was the selected model for both BERTScore and MoverScore, as it was the most powerful Portuguese BERT model readily available. As for BARTScore, there are far fewer models designed for Portuguese. We opted for BART Base Portuguese[2], as it is pre-trained in Portuguese.

In addition to standard BLEU, BLEU-1 and BLEU-2, which use unigrams and bigrams, respectively, were analyzed. As for ROUGE, all ROUGE-1, ROUGE-2, and

[1]Dataset: `huggingface.co/datasets/arubenruben/cnn_dailymail_azure_pt_pt`
[2]Model: `https://huggingface.co/adalbertojunior/bart-base-portuguese`

ROUGE-L were used. For BERTScore, precision, recall, and f1-score values were provided. For more information about each metric, refer to our Github repository[3].

4.3. Instance Selection

The selection of instances for human evaluation was performed at random to avoid bias. For each task, a sample of 40 distinct instances was selected and presented to participants in the human evaluation, as described in the next section.

4.4. Human Evaluation

We used human evaluations to assess metric performance, correlating each score obtained through an automated metric with the scores from the human judgment. During data collection, the four quality dimensions used by SummEval [Fabbri et al. 2021] were employed as to better encompass output quality.

1. Coherence: General quality of the text.
2. Consistency: Alignment between the output and the source text.
3. Fluency: Quality of the generated sentences.
4. Relevance: Key points selected for the text.

However, in the scope of this work, the fluency dimension was replaced by a naturalness one. This change was made because the human evaluations were not conducted by specialists, which impairs the precise analysis of text fluency, as a great amount of linguistic knowledge is required for its evaluation. Naturalness was chosen for its similarity and relative compatibility with fluency, while being easier for laypeople to evaluate.

There are no consolidated quality dimensions designed for the QA evaluation. For this reason, we adopted the same ones that were used for Summarization. We understand that not all dimensions translate directly to this task, but we needed a baseline for the evaluations and considered that maintaining a standard would be a better approach than proposing a distinction without the specific knowledge of an expert.

Data Collection. Human evaluations were collected through two online forms, each focused on one of the tasks. The 40 outputs of each task were divided into two groups and then presented to eight people, leaving four judges for each of the 80 generated texts. The judges were asked to evaluate each of the outputs in the four aforementioned quality dimensions by scoring them on a Likert scale from 1 to 5. To calculate the final score of an output, a simple average of the collected scores was performed.

Participant Profiles. The eight participants were all Computer Science students, aged 21 to 30, and of mixed gender. This makes them knowledgeable in the general functioning of LLMs and familiar with the structure of online forms. However, as the judges were not Linguistic specialists, some nuances may have been lost during data collection.

4.5. Correlation Analysis

To evaluate the alignment between the automated metrics and human evaluation, we adopted both Pearson and Spearman correlations, as they are suitable for continuous data and are widely employed by the literature. Thus, being reliable tools for measuring the

[3]Github: `https://github.com/eduardofae/OET-Metric-Comparison-PT`

correspondence between the metric scores and the collected human data. The discretization of the values adopted for the analysis of this work followed [Schober et al. 2018]. To analyze inter-rater agreement, we calculated Krippendorff's α [Krippendorff 2011] between reviewers, both following the analyzed dimensions and aggregating all scores for each annotator, allowing for dimension-specific and global inter-rater agreement analysis.

5. Results

Table 2 presents the average scores obtained by each metric in each task. The final four rows correspond to the average scores assigned by human evaluation (HE), divided into the four aforementioned quality dimensions. Furthermore, as a means of inter-rater agreement analysis, Krippendorff's Alpha [Krippendorff 2011] between judges is also provided.

Table 2. Average value assigned by each form of evaluation for each task.

	Summarization		QA	
Metric	Avg Score	Krippendorff's α	Avg Score	Krippendorff's α
ROUGE-1	0.22	-	0.34	-
ROUGE-2	0.05	-	0.22	-
ROUGE-L	0.19	-	0.33	-
BLEU	0.01	-	0.18	-
BLEU-1	0.23	-	0.34	-
BLEU-2	0.09	-	0.25	-
METEOR	0.18	-	0.41	-
BERTScore-p	0.66	-	0.75	-
BERTScore-r	0.65	-	0.76	-
BERTScore-f1	0.65	-	0.76	-
BARTScore	-4.25	-	-5.48	-
MoverScore	0.54	-	0.63	-
Consistency (HE)	2.39	0.46	3.44	0.56
Naturalness (HE)	3.58	0.44	4.06	0.43
Relevance (HE)	2.46	0.37	3.48	0.58
Coherence (HE)	2.59	0.37	3.55	0.56

The results indicate substantially lower human evaluation scores for consistency, relevance, and coherence in the summarization task when compared to QA. This behavior may be associated with differences in task complexity, dataset quality, or model capability. However, as this work focuses on the effectiveness of evaluation metrics rather than model performance itself, lower-performing models may even be advantageous, as they produce a wider diversity of both high and low-quality outputs, enabling a better view of metric behavior across different quality levels.

When evaluating the performance of the automatic metrics, although average scores may provide an overview of their raw evaluations, they are insufficient to determine which metrics better reflect human judgment, as each metric operates under distinct scoring scales and evaluation principles. Therefore, correlation analyses are required for a more reliable assessment of the alignment between each automatic metric and human evaluation. These analyses are presented in the following subsections, with results for each task being provided individually.

When it comes to inter-rater agreement, we see relatively low α values for both tasks, with combined values of 0.47 for Summarization and 0.54 for QA. This is a consequence of the open-ended nature of both tasks, as well as the reliance on non-specialized annotators. QA, the more objective of the two tasks, consistently yields higher values than Summarization, which is the broader task. Furthermore, the Naturalness dimension remained constant across tasks, as it is a task-independent dimension, while still exhibiting low values, as it is nevertheless highly subjective. The α values obtained for summarization closely match the ones reported by [Fabbri et al. 2021] from crowd-source workers, signaling the importance of expert reviewers.

5.1. Summarization

Table 3 presents Pearson (ρ) and Spearman (ρ_s) correlation values between each metric and human judgment for the summarization task. The results show relatively weak correlations between metrics and human evaluation. Contrary to our expectations, PLM-based metrics did not outperform n-gram-based ones, with BARTScore [Yuan et al. 2021] particularly underperforming. However, the metric that achieved the highest overall alignment was still a PLM-based one, being it MoverScore [Zhao et al. 2019].

Table 3. Pearson (ρ) and Spearman (ρ_s) correlations between human perception and scores of automatic metrics for the Summarization task.

	Consistency		Naturalness		Relevance		Coherence	
Metric	ρ	ρ_s	ρ	ρ_s	ρ	ρ_s	ρ	ρ_s
ROUGE-1	0.21	0.19	0.29	0.24	0.29	0.27	0.24	0.21
ROUGE-2	0.18	0.07	0.27	0.18	0.23	0.16	0.24	0.13
ROUGE-L	0.28	0.24	0.36	0.31	**0.38**	**0.35**	0.35	0.32
BLEU	0.28	0.23	0.35	0.33	0.33	0.34	0.35	0.33
BLEU-1	0.31	0.28	0.28	0.22	0.28	0.23	0.27	0.22
BLEU-2	0.27	0.22	0.27	0.20	0.25	0.22	0.27	0.20
METEOR	0.32	0.32	**0.37**	**0.35**	0.27	0.31	0.33	0.30
BERTScore-p	0.26	0.21	**0.37**	0.30	0.29	0.23	0.36	0.26
BERTScore-r	0.31	0.24	0.32	0.26	0.29	0.23	0.32	0.22
BERTScore-f1	0.30	0.22	**0.37**	0.30	0.30	0.24	0.36	0.25
BARTScore	0.05	0.04	0.30	0.34	0.15	0.16	0.24	0.25
MoverScore	**0.38**	**0.37**	0.31	0.24	0.34	0.34	**0.39**	**0.36**

Similar findings were previously reported by SummEval [Fabbri et al. 2021] and related studies, reinforcing the difficulty of intrinsically evaluating abstractive summarization systems. Still, the correlations obtained in this work were generally lower than those reported by previous studies. Which may be associated with limitations in the dataset quality, as automatic metrics fundamentally depend on reliable reference texts.

Table 4 shows limitations of the automatic metrics observed in the summarization task. In the first example, the generated summary conveys the central meaning of the source content, resulting in a high human score, yet all metrics, especially n-gram-based ones, assigned relatively low scores. In contrast, the second example entirely cuts out important details, such as the death of a band member and how this is the band's last tour, which leads to a low human score. However, it was better evaluated by all automatic

metrics, except MoverScore, which still scored them close to equally. Indicating that metrics may capture topical similarity more effectively than factual faithfulness.

Table 4. Examples illustrating failures of automatic metrics in summarization.

Example 1 (High-quality Summary)	**Example 2 (Low-quality Summary)**
Summary: Iraqi forces are being sent to central Iraq to help fight the extremist group calling itself the Islamic State (IS). **Human Score:** 4.5/5	**Summary:** The band Twisted Sister disclosed on Wednesday that they'll be leaving the band in the next two months, after a three-year tour. **Human Score:** 1.6/5

5.2. Question Answering

Unlike the behavior observed in the summarization task, the evaluated metrics achieved marginally higher correlation scores in the QA task, as seen in Table 5. This behavior may be explained by the task's underlying characteristics, which, even following an abstractive approach, remain closer to closed-ended settings than summarization, as the outputs are typically shorter and frequently contain terms directly extracted from the source context. Consequently, metrics based on text comparison tend to perform better, as reference texts and outputs are more inclined to share a similar structure and terms.

Table 5. Pearson (ρ) and Spearman (ρ_s) correlations between human perception and scores of automatic metrics for the QA task.

	Consistency		Naturalness		Relevance		Coherence	
Metric	ρ	ρ_s	ρ	ρ_s	ρ	ρ_s	ρ	ρ_s
ROUGE-1	0.58	0.59	0.45	0.54	0.56	0.56	0.55	0.54
ROUGE-2	0.50	0.57	0.33	0.52	0.51	0.53	0.49	0.54
ROUGE-L	0.56	0.58	0.44	0.53	0.55	0.55	0.54	0.52
BLEU	0.43	0.49	0.24	0.31	0.44	0.48	0.43	0.48
BLEU-1	0.51	0.54	0.36	0.50	0.50	0.51	0.49	0.48
BLEU-2	0.53	0.59	0.37	0.55	0.53	0.56	0.52	0.57
METEOR	0.64	0.68	**0.53**	**0.65**	0.63	0.65	0.62	0.61
BERTScore-p	0.63	0.65	0.42	0.47	0.62	0.64	0.60	0.58
BERTScore-r	**0.70**	**0.76**	0.51	0.62	**0.69**	**0.74**	**0.67**	**0.71**
BERTScore-f1	0.68	0.73	0.47	0.58	0.67	0.71	0.65	0.67
BARTScore	0.47	0.41	0.35	0.39	0.47	0.39	0.47	0.41
MoverScore	0.47	0.67	0.28	0.51	0.48	0.65	0.46	0.61

Despite the comparatively better performances, metrics only achieved an overall medium correlation, with BERTScore [Zhang et al. 2019] scraping the lower bounds of a high correlation in some variant/dimension pairs. This result reinforces the limitations of current intrinsic evaluation methods, even in tasks with shorter and more lexically constrained outputs. Similar findings were also reported by [Junqueira and Moreira 2026], which demonstrated that widely adopted automatic metrics remain weakly correlated with human judgments in the QA task.

Table 6 highlights limitations of automatic metrics observed in the QA tasks. In the first example, the generated answer is semantically equivalent to the reference, leading

to high human scores. Contrarily, in the second example, the generated answer is factually incorrect and only loosely related to the reference, resulting in a low human score. However, most metrics, with the sole exception of BERTScore and MoverScore, attributed higher scores to the second instance. These examples demonstrate how current automatic metrics may fail to accurately capture semantic equivalence and answer validity.

Table 6. Examples illustrating failures of automatic metrics in QA.

Example 3 (High-quality Answer)	**Example 4 (Low-quality Answer)**
Question: What did the Happy Hunter discover when he pulled the line for the last time?	**Question:** What strange thing did the prince do?
Answer: The fishing hook was lost.	**Answer:** He was carrying a dagger in his belt.
Reference: He discovered he lost the fishing hook without knowing when he dropped it.	**Reference:** He didn't use the dagger during the fight.
Human Score: 4.7/5	**Human Score:** 1.5/5

5.3. Discussion

The results reinforce the need for the development of evaluation metrics that better reflect human perception. Although PLM-based metrics achieved slightly higher correlations with human judgments, especially in QA, the observed correlations maintained medium to low values across all evaluated metrics, with outliers scraping the high correlation bound. Additionally, results suggest that n-gram-based metrics may be capable of achieving competitive or even superior performances.

Among the evaluated PLM-based approaches, BARTScore [Yuan et al. 2021] presented notably lower performance than expected. A possible explanation is the limited availability of large BART models pre-trained in Portuguese, since PLM-based metrics strongly depend on the quality of the underlying model.

Among n-gram-based strategies, METEOR [Banerjee and Lavie 2005] showed the highest alignment with human judgment, reinforcing the benefits of a synonym dictionary. While ROUGE-1 and ROUGE-2 [Lin 2004] underperformed in the Summarization task, highlighting their sensitivity to variations in lexical choices and syntactic structures.

The results also highlight the scarcity of well-structured and carefully curated datasets in Portuguese, forcing studies to rely on translated or adapted datasets, introducing additional sources of noise and inconsistency. Since automatic metrics fundamentally depend on reference texts, problems such as translation errors, cultural mismatches, and poorly formulated references directly compromise the reliability of intrinsic evaluation.

6. Conclusion

This work investigated the correlation between automatic evaluation metrics and human judgments in open-ended tasks, with a particular focus on summarization and QA in Portuguese. The obtained results demonstrate that widely adopted automatic metrics still present medium to low correlations with human evaluation. The study also identified important limitations, including the scarcity of high-quality datasets and language models for Portuguese. These factors negatively impact both output generation and the reliability of automatic metrics. Overall, the results reinforce the need for more robust evaluation methods capable of better reflecting human judgment.

Limitations

While this study provides a unified multi-task methodology for analyzing automatic evaluation metrics in Portuguese, several limitations must be acknowledged.

Sample Size and Scale of Human Evaluation. Due to resource constraints, the human-evaluation benchmark was limited to 40 instances per task. Expanding the evaluation to a larger and more diverse set of models and datasets would provide a more comprehensive assessment of metric stability across different quality levels.

Non-Expert Evaluator Profiles. The eight human judges selected for this study were Computer Science students. Although they are knowledgeable in the general functioning of LLMs, they lack formal linguistic training, meaning some subtle nuances may have been lost during data collection.

Dependence on Machine-Translated Datasets. This study relied on machine-translated benchmarks, which inherently introduce specific artifacts, translation errors, and semantic distortions. Since automatic metrics are bound to the quality of their reference texts, these limitations directly constrain the reliability of the calculated correlations.

Constraints of Under-Resourced Language Models. The performance of PLM-based metrics is highly dependent on the quality of the underlying models. Any non-English language, such as Portuguese, suffers from the scarcity of large, natively pre-trained models. This limitation was particularly evident in BARTScore's results, which yielded correlation scores that were drastically lower than expected.

Cross-Task Adaptation of Quality Dimensions. In the absence of firmly consolidated, standardized evaluation dimensions designed specifically for QA, this work mirrored the four dimensions used in summarization. Although this choice maintained the baseline across both tasks, certain dimensions may not perfectly translate to the QA task.

Ethics Statement

All human evaluations were collected with the participants' written consent; they were fully informed about the study's purpose, their participation was entirely voluntary, and they were allowed to withdraw at any time without penalty. No personal data was collected from the participants, apart from which group they belonged to. The participants were all closely related to the authors, but were advised that trying to inflate the results would only lead to a worse performance for the metrics.

AI Usage

Generative AI was used as an auxiliary tool during research and revision of this paper. For research, the generative models' ability to conduct in-depth research was used to scan the literature and find works that support specific claims or findings. Generative AI was not the primary source for research, being used in limited cases and always with caution, as we understand information obtained through this means is often unreliable and may be impacted by hallucinations. For revision, Generative AI was used to correct grammar, suggest phrasing, and analyze weaknesses.

Acknowledgments

This work has been partially funded by CNPq and Capes Finance Code 001.

References

Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures*, pages 65–72.

Deutsch, D., Dror, R., and Roth, D. (2021). A statistical analysis of summarization evaluation metrics using resampling methods. *Transactions of the Association for Computational Linguistics*, 9:1132–1146.

Deutsch, D., Dror, R., and Roth, D. (2022). Re-examining system-level correlations of automatic summarization evaluation metrics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 6038–6052.

Fabbri, A. R., Kryściński, W., McCann, B., et al. (2021). SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Farea, A., Yang, Z., Duong, K., et al. (2025). Evaluation of question answering systems: Complexity of judging a natural language. *ACM Comput. Surv.*

Hasan, T., Bhattacharjee, A., Islam, M. S., et al. (2021). Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. *arXiv preprint arXiv:2106.13822*.

Junqueira, J. d. R. and Moreira, V. P. (2026). The inadequacy of automatic evaluation metrics in question answering: A case-study in Portuguese. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 551–561, Salvador, Brazil. Association for Computational Linguistics.

Koehn, P. and Monz, C. (2006). Manual and automatic evaluation of machine translation between European languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121.

Krippendorff, K. (2011). Computing krippendorff's alpha-reliability.

Krishna, K., Roy, A., and Iyyer, M. (2021). Hurdles to progress in long-form question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the ACL*, pages 4940–4957.

Leite, B., Osório, T. F., and Cardoso, H. L. (2024). Fairytaleqa translated: Enabling educational question and answer generation in less-resourced languages. In *Technology Enhanced Learning for Inclusive and Equitable Quality Education*, pages 222–236.

Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Liu, C.-W., Lowe, R., Serban, I., et al. (2016). How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response

generation. In Su, J., Duh, K., and Carreras, X., editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Liu, Y., Iter, D., Xu, Y., et al. (2023). G-eval: NLG evaluation using gpt-4 with better human alignment. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Nallapati, R., Zhou, B., dos Santos, C., et al. (2016). Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290.

NIST, N. (2006). Nist 2006 machine translation evaluation official results.

Paiola, P. H., de Rosa, G. H., and Papa, J. P. (2022). Deep learning-based abstractive summarization for brazilian portuguese texts. In *BRACIS 2022: Intelligent Systems*, pages 479–493.

Papineni, K., Roukos, S., Ward, T., et al. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Rajpurkar, P., Zhang, J., Lopyrev, K., et al. (2016). Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Schober, P., Boer, C., and Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, 126(5):1763–1768.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020*, page 403–417.

Team, G., Kamath, A., Ferret, J., et al. (2025). Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Xu, Y., Wang, D., Yu, M., et al. (2022). Fantastic questions and where to find them: Fairytaleqa – an authentic dataset for narrative comprehension.

Yuan, W., Neubig, G., and Liu, P. (2021). Bartscore: Evaluating generated text as text generation. *Advances in neural information processing systems*, 34:27263–27277.

Zhang, T., Kishore, V., Wu, F., et al. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhao, W., Peyrard, M., Liu, F., et al. (2019). Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv:1902.02622*.

DANTE – A Treebank of Tweets in Brazilian Portuguese

Ariani Di-Felippo[1,2], **Norton Trevisan Roman**[3]

[1]Language and Literature Department (DL/UFSCar), São Carlos – Brazil

[2]Interinstitutional Center for Computational Linguistics (NILC/USP) São Carlos – Brazil

[3]School of Arts, Sciences and Humanities (EACH/USP), São Paulo – Brazil

ariani@ufscar.br, norton@usp.br

Abstract. *In this article, we present the current state of DANTE (Dependency-ANalysed corpora of TwEets), a treebank focusing on tweets (currently X posts) written in Brazilian Portuguese, featuring four subcorpora varying in size, domain, and level of annotation under the Universal Dependencies framework. This article aims to inform the Natural Language Processing and Linguistic communities about currently available corpora, as well as what to expect from the treebank in the near future, while also bringing together the tools and language resources developed within the DANTE project.*

1. Introduction

When it comes to user-generated content[1] (UGC), the existence of annotated treebanks[2] becomes an invaluable resource, as they enable the modelling of highly variable, noisy, and domain-specific language typical of social media [Sanguinetti et al. 2023]. Among the theoretical frameworks for the construction of treebanks, the Universal Dependencies (UD) project [de Marneffe et al. 2021] stands out as a widely adopted international initiative that provides a unified scheme for morphosyntactic and syntactic annotations across languages, with over 150 languages available in more than 200 corpora.

Among the resources available in the UD repository, Portuguese is represented by seven corpora, comprising five general-domain resources (Bosque [Rademaker et al. 2017], GSD[3], PUD [Zeman 2017], CINTIL [Branco et al. 2022], and Porttinari-base [Duran et al. 2023]), one domain-specific corpus in the oil and gas domain (PetroGold [Souza et al. 2021]), and one UGC corpus (DANTEStocks [Di-Felippo and Roman 2025]). Porttinari-base is a carefully revised corpus designed to serve as a gold standard within the multigenre treebank Porttinari ("PORTuguese Treebank")[4]. In addition to this component, the treebank also includes Porttinari-check, a smaller corpus used as a testbed for comparing manual and automatic annotations, and Porttinari-automatic, a large corpus annotated automatically.

Similarly to Porttinari-base, DANTEStocks is part of a larger treebank – DANTE (Dependency-ANalysed corpora of TwEets), also integrating Porttinari. This resource is a pioneering tweet-based treebank, currently comprising data from two domains, along with

[1]UGC is any form of content, such as images, videos, text and audio, that has been posted on online platforms such as social media, review pages, blogs and wikis [Krumm et al. 2008].

[2]A corpus in which every sentence is annotated with a parse tree [Jurafsky and Martin 2026].

[3]https://universaldependencies.org/treebanks/pt_gsd/index.html

[4]https://sites.google.com/icmc.usp.br/poetisa/porttinari-3-0

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 139–149
October 19-22, 2026 ©2026 Association for Computational Linguistics

UD annotation. The treebank is organised into four subcorpora with different characteristics and stages of development: DANTEStocks, DANTEStocks-Large, DANTEShots and DANTEShots-Extended. The first two were drawn from the stock market domain, while the latter two focus on posts by Brazilian politicians regarding COVID-19 vaccination. This article describes then the current state of DANTE, outlining what is already available and what is to be expected for the near future. In addition, it situates DANTEStocks – the only subcorpus with morphosyntactic (Part-of-Speech – PoS) and syntactic gold-standard annotations under the UD framework – within the broader set of UGC resources available in the UD project.

The rest of this article is organised as follows. Section 2 provides an overview of the currently available UGC corpora in the UD project, discussing the role of Portuguese within this landscape, with DANTEStocks as its representative resource. Section 3, in turn, focuses on the DANTE treebank, describing its structure, subcorpora, and annotation layers. In Section 4, some NLP tools and language resources developed based on DANTE's annotations are listed, with Section 5 presenting our final considerations, also giving a brief overview of our future plans for extending each subcorpus.

2. Treebanks featuring UGC in the UD project

In a recent overview of the UGC treebanks available in the literature, [Sanguinetti et al. 2023] have found 30 resources published between 2011 and 2020. This study was based on a semi-systematic Search on Google Scholar for open-access papers and complemented by additional known resources. Unlike this overview, we focus exclusively on corpora standardised according to the UD annotation scheme and integrated into the official repository of the UD project[5], as summarised in Table 1.

The table reports the number (#) and genre of each corpus, their data sources as well as the total number of annotated sentences and tokens. Currently, the UD repository comprises 22 corpora covering 18 languages. Based on this table, the amount of data varies considerably across languages, from South Levantine Arabic, with 789 tokens, to Russian, with nearly 1,8 million tokens. Being represented by DANTEStocks, with 4,042 tweets from the stock market, Portuguese ranks 10th among languages. With 77,576 tokens, it occupies a mid-range coverage in comparison to the other UGC corpora of the UD initiative.

Six out of the 22 corpora outlined in the table are either entirely (TwittIrish, PoSTWITA, TWITTIRO, DANTEStocks, and TueCL) or partially (OOD) composed of Twitter data. This suggests a moderate prevalence of Twitter-based resources, with the platform remaining as a widely used source for short, informal text. Within the subset of corpora exclusively based on Twitter data, Portuguese ranks 2nd in terms of size, with 77,576 tokens. It is surpassed only by Italian, which collectively add up to 147,718 tokens (PoSTWITA and TWITTIRO). Despite the limited resource diversity for Portuguese, its single available corpus exhibits a relatively high token density within this subdomain. Although DANTEStocks provides a non-negligible volume of annotated data, its linguistic coverage remains nevertheless restricted to a narrow communicative setting (to wit, the stock market domain).

[5] https://universaldependencies.org/

Table 1. Overview of UGC corpora in the UD project.

Language	#	Corpus	Genre	UGC source	Sents.	Tokens
Armenian	1	ArmTDP	Various	Web, social	5,527	103,846
Bavarian	1	MaiBaam	Various	Social	1,070	14,678
Belarusian	1	HSE	Various	Web, social	25,231	305,109
Czech	1	PDTC	Various	Web, social	5,527	103,846
English	4	GUM	Various	Reddit	13,263	229,851
		EWT	Various	Web, social	16,662	251,489
		GENTLE	Various	Web, social	1,334	17,619
		GUMReddit	Various	Reddit	859	15,958
Estonian	1	EWT	Various	Web, social	7,190	60,583
Finish	1	OOD	Social	Various (Twitter)	2,122	19,364
Irish	1	TwittIrish	Social	Twitter	2,596	47,790
Italian	2	PoSTWITA	Social	Twitter	6,712	119,334
		TWITTIRO	Social	Twitter	1,424	28,384
Manx	1	Cadham	Various	Web, social	2,336	18,580
Persian	1	Seraji	Various	Web, social	5,997	151,627
Polish	1	LFG	Various	Web, social	17,246	130,967
Portuguese	1	DANTEStocks	Social	Twitter	4,042	77,576
Romanian	1	TueCL	Social	Twitter	270	4,417
Russian	1	Taiga	Various	Social media	121,967	1,758,937
South Levantine Arabic	1	MADAR	Social	Spoken	100	789
Ukrainian	1	IU	Various	Social	7,092	122,701
Western Armenian	1	ArmTDP	Various	Social	6,644	121,432
Total	22				255,211	3,704,877

3. DANTE's Current State

As for the present moment, DANTE comprises four corpora (Table 2), with its first corpus - DANTEStocks - being the richest one in terms of annotations added, also being the only fully available to the community[6]. Altogether, the four corpora feature more than 146,000 tweets, with over 137,000 of them already annotated with some UD layer, either manually or automatically, along with some semantic information, such as named entities, emotions, sentiment and stance, depending on the corpus. In what follows, we present each corpus in more detail, along with their current development stage.

3.1. DANTEStocks

DANTEStocks is the first tweebank in Brazilian Portuguese annotated within the UD framework. The corpus consists of tweets from the stock market domain, collected in 2014 based on mentions to Ibovespa-listed companies, and reflects typical Twitter constraints until 2017 (*e.g.* the 140-character limit). The original resource that gave rise to DANTEStocks was already annotated with emotions following Plutchik's model [Plutchik 2001]. DANTEStocks extends this original resource via standoff annotation, adding morphological (PoS, lemmas and features) and dependency layers following UD guidelines, all curated and validated by experts. The UD-annotation of the corpus

[6]Within the POeTiSA project: `https://sites.google.com/icmc.usp.br/poetisa`

Table 2. UGC corpora in the DANTE treebank.

Subcorpus	Domain	Annotation	Status	Tweets	Tokens
DANTEStocks	Stock market	UD (complete) Emotion Named entity	Gold-standard Gold-standard Gold-standard	4,042	77,576
DANTEStocks-Large	Stock market	UD PoS Emotion	Automatic Automatic	126,579	3,932,604
DANTEShots	Covid-19 vaccines	UD PoS Stance	Automatic Gold-standard	7,056	296,859
DANTEShots-Extended	Covid-19 vaccines	Stance Sentiment	Gold-standard Gold-standard	9,045	
Total				146,722	4,307,039

adopts tweets as the basic unit of analysis, avoiding sentence-level segmentation and normalization [Di-Felippo et al. 2021].

Tokenization was carried out in a semi-automatic way, by using DANTE's Tokenizer [Silva et al. 2021], which is a rule-based system adapted to UGC phenomena. At the morphological layer, the PoS tags were generated using UDPipe 2 [Straka et al. 2016] pre-trained on the Bosque corpus [Rademaker et al. 2017], and iteratively refined through human review and retraining [Di-Felippo et al. 2022, Felippo et al. 2023]. Lemmas and morphological features, as described in [Di-Felippo and Roman 2025], were added semi-automatically using PortiLexicon-UD [Lopes et al. 2022]. Syntactic annotation relied on a hybrid strategy, starting from a manually revised gold-standard subcorpus[7] and incrementally training a parser[8] with iterative human validation [Di-Felippo et al. 2024a, Barbosa 2024], supported by guidelines specifically designed for Portuguese [Duran 2021] and for Twitter data [Di-Felippo et al. 2024b].

The corpus was later annotated with Named Entities [Zerbinati et al. 2024] following the Second HAREM's taxonomy [Mota and Santos 2008]. This annotation was then revised and augmented [Piai et al. 2025], by including a fine-grained, linguistically motivated annotation, tailored to the financial domain of tweets in Portuguese, extending HAREM's traditional categories to better capture entities such as stock market instruments (e.g., tickers), financial indicators, and other market-specific expressions. The annotation was carried out manually by a single annotator, following detailed guidelines and iterative refinement to ensure consistency, including explicit rules for delimiting entity boundaries in noisy data (*e.g.* orthographic variation and truncation) and context-based disambiguation strategies to ensure consistent classification of financial entities.

Recently, [Ceregatto and Felippo 2025] have extended the corpus with a subset of 1,128 distinct tweets manually annotated according to the Abstract Meaning Representation (AMR) framework [Banarescu et al. 2013, Banarescu et al. 2019]. Additionally, [Barbosa 2024] provided argument structure for 1,756 instances (and 1,218 tweets) of

[7]Obtained by automatically annotating 1,000 tweets with UDPipe 2 [Straka et al. 2016] trained on Bosque [Rademaker et al. 2017], followed by manual revision.

[8]Our approach was to begin with the Stanza [Qi et al. 2020] base architecture, fine-tuned on Porttinari-base and our reference subcorpus.

146 predicate nouns[9] in DANTEStock, following NomBank [Meyers 2007] for English, which is a parallel project to PropBank [Palmer et al. 2005, Pradhan et al. 2022].

3.2. DANTEStocks-Large

DANTStocks-Large was collected following the same methodology as the corpus that originated DANTEStocks (described in [da Silva et al. 2020]). Beyond collecting a considerably larger set of tweets, the corpus' goal was also to add date and time information to them, an information that was missing in DANTStocks. In total, over 126 thousand tweets were collected from August, 2022 to September, 2023. These were then automatically annotated with emotions, following the Plutchik's Wheel of Emotions model [Plutchik 2001]. To do so, we relied on classifiers developed and tested in the corpus by [da Silva et al. 2020], as presented in [de Carvalho and Roman 2025]. Besides emotions, we have also adapted the tagger presented in [Silva et al. 2023] and used it to automatically tokenise and annotate the corpus with PoS under UD's framework. The resulting corpus comprises then over 126 thousand tweets annotated with emotions, with almost 4 million tokens automatically annotated with UD-PoS. Before officially presenting the corpus to the community, however, we intend to carry out a human adjudication of the annotations made, both for emotion and PoS.

3.3. DANTEShots

DANTEShots builds on an existing corpus of tweets related to COVID-19 vaccines [Barberia et al. 2023], as posted by political candidates during 2020 and 2021, annotated with stance towards vaccination. This corpus was then automatically annotated with UD-PoS tags, by using an adapted version of the tagger presented in [Silva et al. 2023] (the same used in DANTEStocks-Large), which resulted in a corpus of 7,056 tweets and 296.859 annotated tokens. To assess the annotation's quality, a sample of 1,000 tweets was randomly[10] taken from the corpus, being currently under review by a set of three human adjudicators. The results of the adjudication will be later used to evaluate the tagger's performance, also giving researchers an estimate of labels' accuracy.

3.4. DANTEShots-Extended

The same research group responsible for the corpus of COVID-19 vaccine posts used as the basis for DANTEShots (*cf.* [Barberia et al. 2023]), later on released a larger corpus, this time with 9,045 tweets annotated for relevance, stance and sentiment towards COVID-19 vaccines and vaccination during the pandemics [Barberia et al. 2025]. This new corpus extends the previous one by applying a new annotation scheme and increasing the corpus' size so as to include posts from 2020 to 2022 by local political elites. Since DANTEShots was based on the smallest version of this base corpus, we also decided to augment it with this new corpus, thereby giving birth to DANTEShots-Extended. This is, however, by far the most embryonic of all corpora within DANTE, and efforts are just in their initial stages to automatically annotate the corpus with PoS and syntactic structure. Once finished, the corpus will allow for the connection between morphosyntax, syntax, stance and sentiment. Although we have no fixed deadline to this task at this moment, we expect it to happen sometime next year.

[9] A predicative noun is one that selects an argument structure (valency), associating semantic participants with its lexical meaning.

[10] From a uniform distribution.

4. NLP Tools and Lexical Resources

DANTEStocks has already led to the development of various NLP tools and resources for Brazilian Portuguese[11]. One of these tools is the DANTE tokenizer [Silva et al. 2021], which is a rule-based system developed for stock-market tweets, built on top of the NLTK TweetTokenizer and extended with linguistically motivated regular expressions. Another tool is its UGC Parser [Di-Felippo et al. 2024a, Barbosa 2024], a UD parsing model derived from Stanza [Qi et al. 2020]. The model was initially trained on a combination of Porttinari-base and a gold-standard reference subcorpus, and subsequently improved through an incremental training strategy in which newly annotated tweet packages were automatically parsed, manually revised, and incorporated into successive training cycles. DANTEStocks' gold-standard syntactic (UD-based dependencies) annotation also supported the development of Genipapo [Di-Felippo et al. 2024c], a multi-genre (tweets, news, and academic texts) parser, achieving better or competitive performance in relation to genre specific trained parsers.

Regarding lexical resources, the semantic annotation of a subset of tweets from DANTEStocks inspired in NomBank gave rise to NounBank.DS [Barbosa and Felippo 2025] – an online repository that provides argument structure for instances of predicate nouns[12]. More recently, a descriptive study on DANTEStocks' vocabulary [Di-Felippo et al. 2026] resulted in a lexicon of orthographic variations, innovative forms, and medium- and domain-dependent tokens, which has already been proved to reduce the out-of-vocabulary rate produced by different embedding models. This vocabulary (LexDANTE) is available for online access and download through a dedicated repository.[13]

5. Final Remarks and Directions for Future Research

This paper outlined our current efforts in building DANTE, a treebank focusing on tweets from different domains, as a part of a larger effort to produce Porttinari, a multi-genre treebank for written texts in Brazilian Portuguese. Connecting morphosyntactic, syntactic and, sometimes, semantic information, at various levels and in different domains, we hope that DANTE can substantially contribute to improve the NLP community's ability to understand and process tweets/posts in Brazilian Portuguese.

Currently, efforts within the DANTE project focus on domain diversification, with the development of corpora from different domains (*e.g.* stock market and vaccination), aiming at their future integration into the UD repository and broader availability to the NLP community. As such, we expect to release DANTEShots by the end of this year, with the automatically generated UD annotations fully revised by human annotators. As a next step, we plan to apply a UGC Parser [Di-Felippo et al. 2024c] to automatically generate dependency annotations, thereby extending the corpus with a syntactic layer. Similar developments are planned for DANTEStocks-Large and DANTEShots-Extended, although DANTEShots-Extended is scheduled for release next year only.

[11] Available from `https://sites.google.com/icmc.usp.br/poetisa`

[12] `https://bryankhelven.github.io/NounBank.DS/`

[13] `https://bryankhelven.github.io/LexDANTE/`

Acknowledgements

This work was carried out at the Center for Artificial Intelligence of the University of São Paulo (C4AI - http://c4ai.inova.usp.br/), with support by the São Paulo Research Foundation (FAPESP grant 2019/07665-4) and by the IBM Corporation. The project was also supported by the Ministry of Science, Technology and Innovation, with resources of Law N. 8.248, of October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published as Residence in TIC 13, DOU 01245.010222/2022-44.

Limitations

When upgrading the SRL component to a higher-quality neural model, we re-evaluate the ablation-based rule analysis under the updated intermediate representation, in order to assess the robustness of the rule hierarchy to changes in semantic role prediction quality.

Ethics Statement

All tweets in DANTE are public, and can be publicly accessed through X's website by using their identification numbers. Still, profile identification was removed from all data sets so as to make it harder to determine personal information of tweet authors.

Disclosure Regarding the Use of AI Tools

AI was used in some parts of the article as a way to check the appropriateness of some grammatical structures used in the text.

References

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N. (2013). Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N. (2019). Abstract meaning representation (amr) 1.2.6 specification. Technical report, AMR Project. Version 1.2.6.

Barberia, L., Schmalz, P., Roman, N. T., Lombard, B., and de Sousa, T. M. (2025). It's about what and how you say it: A corpus with stance and sentiment annotation for covid-19 vaccines posts on x/twitter by brazilian political elites. In Hämäläinen, M., Öhman, E., Bizzoni, Y., Miyagawa, S., and Alnajjar, K., editors, *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities (NLP4DH 2025)*, pages 365–376, Albuquerque, USA. Association for Computational Linguistics.

Barberia, L. G., de Santana Schmalz, P. H., and Roman, N. T. (2023). When tweets get viral - a deep learning approach for stance analysis of covid-19 vaccines tweets by brazilian political elites. In *Proceedings of the 14th Symposium in Information and Human Language Technology (STIL 2023)*, Belo Horizonte, MG - Brazil.

Barbosa, B. and Felippo, A. D. (2025). Nounbank.ds: a lexical repository of nominal frames from stock market tweets in brazilian portuguese. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 29–41, Porto Alegre, RS, Brasil. SBC.

Barbosa, B. K. S. (2024). Descrição sintático-semântica de nomes predicadores em tweets do mercado financeiro em português. Dissertação (mestrado em linguística), Universidade Federal de São Carlos, São Carlos, SP.

Branco, A., Silva, J. R., Gomes, L., and António Rodrigues, J. (2022). Universal grammatical dependencies for Portuguese with CINTIL data, LX processing and CLARIN support. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5617–5626, Marseille, France. European Language Resources Association.

Ceregatto, G. and Felippo, A. D. (2025). Dantestocks-amr em construção: Avanços e desafios na anotação semântica de tweets financeiros. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 608–617, Porto Alegre, RS, Brasil. SBC.

da Silva, F. J. V., Roman, N. T., and Carvalho, A. M. B. R. (2020). Stock market tweets annotated with emotions. *Corpora*, 15(3):343–354. Online ISSN: 1755-1676.

de Carvalho, W. P. and Roman, N. T. (2025). Classifying emotions in tweets from the financial market: A bert-based approach. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing (RANLP 2025)*, pages 218–226, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Di-Felippo, A., das Graças Nunes, M., and Barbosa, B. (2024a). A dependency treebank of tweets in brazilian portuguese: Syntactic annotation issues and approach. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 192–201, Porto Alegre, RS, Brasil. SBC.

Di-Felippo, A., Nunes, M. d. G. V., and Barbosa, B. K. d. S. (2024b). Diretrizes de anotação de relações de dependência em tweets do mercado financeiro. Technical Report n. 446, Relatório Técnico – ICMC, USP, São Carlos.

Di-Felippo, A., Postali, C., Ceregatto, G., Gazana, L., Silva, E., Roman, N., and Pardo, T. (2021). Descrição preliminar do corpus DANTEStocks: diretrizes de segmentação para anotação segundo Universal Dependencies. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 335–343, Porto Alegre, RS, Brasil. SBC.

Di-Felippo, A., Postali, C., Ceregatto, G., Gazana, L. S., and Roman, N. T. (2022). Diretrizes de anotação de pos tags em tweets do mercado financeiro: Orientações para anotação em língua portuguesa segundo a abordagem universal dependencies. Technical Report n. 438, Relatório Técnico – ICMC, USP, São Carlos-SP.

Di-Felippo, A. and Roman, N. T. (2025). Dantestocks: A multi-layered annotated corpus of stock market tweets for brazilian portuguese. *Revista Brasileira de Linguística Aplicada*, 25(1).

Di-Felippo, A., Roman, N. T., Barbosa, B. K. S., Oliveira, G. P. d., and Scandarolli, C. L. (2026). Lexical and orthographic variation in Portuguese financial tweets: Annotation, analysis, and implications for embedding models. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 628–637, Salvador, Brazil. Association for Computational Linguistics.

Di-Felippo, A., Roman, N. T., da Silva Barbosa, B. K., and Pardo, T. A. S. (2024c). Genipapo – a multigenre dependency parser for brazilian portuguese. In *Proceedings of the 15th Brazilian Symposium in Information and Human Language Technology (STIL 2024)*, pages 257–266, Belém - PA, Brazil.

Duran, M., Lopes, L., Nunes, M. d. G. V., and Pardo, T. A. S. (2023). The dawn of the Porttinari multigenre treebank: introducing its journalistic portion. In *Proceedings of the XIV Brazilian Symposium in Information and Human Language Technology (STIL)*, pages 115–124, Porto Alegre, RS, Brasil. SBC.

Duran, M. S. (2021). Manual de anotação de relações de dependência: orientações para anotação de relações de dependência sintática em língua portuguesa, seguindo as diretrizes da abordagem universal dependencies (ud). Technical Report n. 435, Relatório Técnico – ICMC, USP, São Carlos.

Felippo, A. D., Roman, N. T., Pardo, T. A. S., and de Moura, L. P. (2023). The dantestocks corpus: an analysis of the distribution of universal dependencies-based part-of-speech tags. *Revista da Abralin*, 22(2):249–271.

Jurafsky, D. and Martin, J. H. (2026). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd edition. Online manuscript released January 6, 2026.

Krumm, J., Davies, N., and Narayanaswami, C. (2008). User-generated content. *IEEE Pervasive Computing*, 7(4):10–11.

Lopes, L., Duran, M., Fernandes, P., and Pardo, T. (2022). PortiLexicon-UD: a Portuguese lexical resource according to Universal Dependencies model. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6635–6643, Marseille, France. European Language Resources Association.

Meyers, A. (2007). Annotation guidelines for nombank: Noun argument structure for propbank. Technical report, NomBank Project, S.l.

Mota, C. and Santos, D. (2008). *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Piai, L., Di-Felippo, A., and Roman, N. T. (2025). Named entities in stock market tweets: A fine-grained and linguistically-motivated annotation. In *Proceedings of the 16th Brazilian Symposium in Information and Human Language Technology: 10th Portuguese Description Journey (JDP 2025)*, pages 654–663, Fortaleza, CE, Brazil.

Plutchik, R. (2001). The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350.

Pradhan, S., Bonn, J., Myers, S., Conger, K., O'Gorman, T., Gung, J., Wright-Bettner, K., and Palmer, M. (2022). Propbank comes of age—larger, smarter, and more diverse. In Nastase, V., Pavlick, E., Pilehvar, M. T., Camacho-Collados, J., and Raganato, A., editors, *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 278–288, Seattle, Washington. Association for Computational Linguistics.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. In Celikyilmaz, A. and Wen, T.-H., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Rademaker, A., Chalub, F., Real, L., Freitas, C., Bick, E., and de Paiva, V. (2017). Universal Dependencies for Portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 197–206, Pisa, Italy. Linköping University Electronic Press.

Sanguinetti, M., Bosco, C., Cassidy, L., and et al. (2023). Treebanking user-generated content: a ud based overview of guidelines, corpora and unified recommendations. *Language Resources Evaluation*, 57:493–544.

Silva, E., Pardo, T., Roman, N., and Fellipo, A. (2021). Universal dependencies for tweets in brazilian portuguese: Tokenization and part of speech tagging. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 434–445, Porto Alegre, RS, Brasil. SBC.

Silva, E. H., Pardo, T. A. S., and Roman, N. T. (2023). Etiquetagem morfossintática multigênero para o português do brasil segundo o modelo "universal dependencies". In *Proceedings of the 14th Symposium in Information and Human Language Technology (STIL 2023)*, Belo Horizonte, MG - Brazil.

Souza, E., Silveira, A., Cavalcanti, T., Castro, M., and Freitas, C. (2021). Petrogold – corpus padrão ouro para o domínio do petróleo. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 29–38, Porto Alegre, RS, Brasil. SBC.

Straka, M., Hajič, J., and Straková, J. (2016). UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA).

Zeman, D. e. a. (2017). CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. ACL.

Zerbinati, M. M., Roman, N. T., and Felippo, A. D. (2024). A corpus of stock market tweets annotated with named entities. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese (PROPOR 2024)*, pages 276–284, Santiago de Compostela, Galicia/Spain.

EpiCorpus-BR: A Named Entity Recognition Corpus for Epidemiological Documents in Portuguese

Christian Freitas, Lilian Berton

[1]Instituto de Ciência e Tecnologia
Universidade Federal de São Paulo (UNIFESP)
São José dos Campos – SP – Brazil

christian.freitas@unifesp.br, lberton@unifesp.br

Abstract. *This paper presents* ***EpiCorpus-BR****, the first annotated corpus for Named Entity Recognition (NER) in Brazilian Portuguese epidemiological surveillance documents, covering the annual SIREVA-SUS reports (2013–2024) and complementary documents from the Brazilian Ministry of Health. The corpus comprises* ***8,314 textual units*** *across 22 documents (6,358 narrative text sentences and 1,956 table rows), with* ***4,159 entity mentions*** *in the silver set. The tagset consists of eight specialized categories (*PATOGENO*,* SOROTIPO*,* FAIXA_ETARIA*,* LOCAL*,* PERIODO_TEMPORAL*,* METODO_LAB*,* METRICA_EPI*,* MANIFESTACAO_CLINICA*), annotated via few-shot prompting with GPT-4o-mini and manually reviewed in a stratified sample of* ***280 units*** *(180 text sentences and 100 table rows) by two independent annotators (κ = 0.76). Strict IOB2 evaluation with seqeval yields a combined micro-F_1 of 0.77.* METRICA_EPI *is absent from narrative text but reaches F_1 = 0.89 on table rows, which shows why the tabular sub-corpus must be included in the evaluation. A BERTimbau baseline fine-tuned on the silver corpus, with the gold held out, reaches micro-F_1 = 0.73, showing that the corpus supports training a dedicated NER model. The corpus and code are publicly available.*

1. Introduction

In Brazil, epidemiological surveillance of invasive diseases is carried out by the Health Surveillance Secretariat (SVS) of the Ministry of Health [Secretaria de Vigilância em Saúde 2024]. It produces a large volume of technical documents: annual reports, bulletins, and advisories. These texts describe, in specialized prose, etiological agents, circulating serotypes, age groups, laboratory methods, and morbidity and mortality metrics. Despite this richness, the content is still out of reach for automatic information extraction. English already has well-established corpora for biomedical Named Entity Recognition (NER), such as i2b2/n2c2 [Uzuner et al. 2011] and NCBI Disease [Doğan et al. 2014], but no comparable resource exists for the epidemiological surveillance domain in Portuguese [Albuquerque et al. 2023].

This work is part of a line of NLP research applied to SIREVA-SUS documents [Freitas et al. 2026, Freitas et al. 2025], introducing NER [Nadeau and Sekine 2007] as a semantic annotation layer. EpiCorpus-BR proposes a tagset of eight specialized categories (PATOGENO, SOROTIPO,

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 150–162
October 19-22, 2026 ©2026 Association for Computational Linguistics

FAIXA_ETARIA, LOCAL, PERIODO_TEMPORAL, METODO_LAB, METRICA_EPI, and MANIFESTACAO_CLINICA), absent from traditional biomedical ontologies [Uzuner et al. 2011, Doğan et al. 2014] and nonexistent in Portuguese-language resources [Albuquerque et al. 2023].

The construction pipeline combines four steps: (i) narrative text extraction via Docling (scenario B2 from [Freitas et al. 2026]) and table row extraction via pdfplumber, (ii) silver annotation via few-shot prompting with GPT-4o-mini [OpenAI 2024], (iii) manual review of a stratified sample to build the gold standard, and (iv) evaluation with seqeval in the strict IOB2 (*Inside–Outside–Beginning*) labelling scheme. The corpus covers 12 annual SIREVA-SUS reports (2013–2024) and 10 complementary documents, totalling 22 documents, 8,314 textual units (6,358 sentences and 1,956 table rows), and 4,159 entity mentions in the silver set.

The contributions of this work are: 1) An NER tagset with eight specialized epidemiological categories for the invasive disease surveillance domain in Brazilian Portuguese (Section 3). 2) A semi-automatic annotation pipeline combining structured extraction via Docling, silver annotation with a large language model (LLM), and manual review, replicable for other specialized domains without available training data (Section 4). 3) To the best of our knowledge, the first public NER corpus for epidemiological surveillance documents in Portuguese, covering 12 years of SIREVA-SUS data and complementary documents from the Ministry of Health (Section 5). 4) A per-category evaluation using seqeval metrics comparing the silver annotation against the gold standard, a second-LLM annotator comparison (GPT-4o-mini vs. Gemini 2.5 Flash), and a supervised BERTimbau baseline trained on the silver corpus that demonstrates the corpus's downstream utility (Section 6).

2. Related Work

Biomedical and clinical NER in Portuguese. The survey by [Albuquerque et al. 2023] maps NER resources for Portuguese, showing that most available systems and corpora focus on journalistic and general-domain text, with sparse coverage for specialized biomedical domains. SemClinBr [Oliveira et al. 2022] is the main semantically annotated clinical corpus for Brazilian Portuguese, covering entities according to Unified Medical Language System (UMLS) semantic types (diseases, medications, and procedures) across 1,000 hospital clinical notes. Although a reference for clinical NLP in Portuguese, the UMLS semantic types do not encompass epidemiological surveillance categories such as serotypes, surveillance laboratory methods, population-level metrics, or temporal surveillance periods. On the modelling side, BioBERTpt [Schneider et al. 2020] adapted BERT to the clinical and biomedical domain in Portuguese and became a reference for clinical NER in the language; however, models like this depend on domain-specific annotated corpora for fine-tuning, a resource still nonexistent for population-level epidemiological surveillance. In English, NCBI Disease [Doğan et al. 2014] and i2b2/n2c2 [Uzuner et al. 2011] set standards for biomedical NER, but both focus on individual patient records in hospital settings, a text type fundamentally different from population-scale surveillance reports, where the central entity is not the patient but the etiological agent and its epidemiological distribution; comparable resources for epidemiological surveillance in Portuguese are nonexistent [Albuquerque et al. 2023].

Domain-refined NER corpora in Portuguese. A recent line of work in the Portuguese NLP community has invested in NER corpora with refined tagsets tailored to specialized domains, showing that generic categories (Person, Organization, Location) are insufficient to capture the relevant entities of each area. Examples include PetroGeoNER, in the oil and gas domain [Moreira et al. 2025]; the linguistically motivated annotation of entities in stock-market tweets [Piai et al. 2025]; CDJUR-BR, in the legal domain [Brito et al. 2023]; and approaches for NER in clinical case reports [Ramos et al. 2026] and for medical entities with compact BERT models [Pontes et al. 2024]. EpiCorpus-BR belongs to this tradition of specialized tagsets, but addresses a domain not yet covered: population-scale surveillance of invasive diseases. Its categories (pathogen, serotype, surveillance laboratory method, epidemiological metric) have no counterpart in the previous resources.

Semi-automatic annotation with language models. The use of large language models (LLMs) for generating silver annotations has gained traction as a cost-effective alternative to full manual annotation [Ding et al. 2024], particularly in scenarios where annotated training data are scarce or unavailable. [Ding et al. 2024] organize LLM-based data augmentation strategies into three learning paradigms (training, fine-tuning, and in-context learning) and identify that annotation quality depends critically on prompt design.

3. Named Entity Tagset

The tagset comprises eight closed categories, designed specifically for the epidemiological surveillance domain for invasive diseases in Brazilian Portuguese. The categories were defined through qualitative analysis of SIREVA-SUS reports and iteration over a prior prototype with a different tagset. Table 1 presents the eight categories with definitions and examples extracted from the reports.

Table 1. SIREVA-SUS-NER tagset categories, with definitions and examples from the documents.

Category	Definition	Examples
PATOGENO	Organism causing invasive disease (bacteria, viruses, fungi).	*S. pneumoniae*, *N. meningitidis*, pneumococo
SOROTIPO	Serological identification: serotype, serogroup, or type.	sorotipo 19A, sorogrupo B, tipo b
FAIXA_ETARIA	Age range used as epidemiological stratification.	menor que 12 meses, 60 anos ou mais
LOCAL	Named geographical location (country, state, region).	Brasil, São Paulo, região Sudeste
PERIODO_ TEMPORAL	Explicit temporal span: years, weeks, or months.	2024, 2013 a 2024, semana epidemiológica 35
METODO_LAB	Named laboratory or diagnostic technique.	cultura, PCR em tempo real, MALDI-TOF
METRICA_EPI	Quantitative measure: the named concept in text, each numeric value in tables.	taxa de incidência (text); 42, 97.4 (table)
MANIFESTACAO_ CLINICA	Named disease, syndrome, or clinical condition.	meningite, pneumonia, doença invasiva

3.1. Design Decisions

Separating PATOGENO and SOROTIPO; removing VACINA and ORG. In a prior prototype, serotype and pathogen were fused into a single category, generating ambiguity in expressions such as *"pneumococo sorotipo 19A"*. Separating them allows independent tracking of agent and serotype distributions across the historical series. Categories for vaccines and institutional organizations were removed: vaccines appear primarily as context for vaccination coverage (captured by METRICA_EPI), and organization names introduce high ambiguity without contributing to the central epidemiological analyses.

Distinguishing METRICA_EPI in text vs. tables. In narrative text, the named concept is annotated (*"taxa de incidência"*, *"cobertura vacinal"*); in table rows, each individual numeric value is annotated, since domain metrics are concentrated in tables rather than prose (see Section 6).

4. Corpus Construction Pipeline

The pipeline comprises four sequential steps, illustrated in Figure 1: (i) extraction of narrative text and table rows from PDFs, (ii) automatic silver annotation via few-shot prompting, (iii) manual review to build the gold standard, and (iv) automatic evaluation with seqeval.

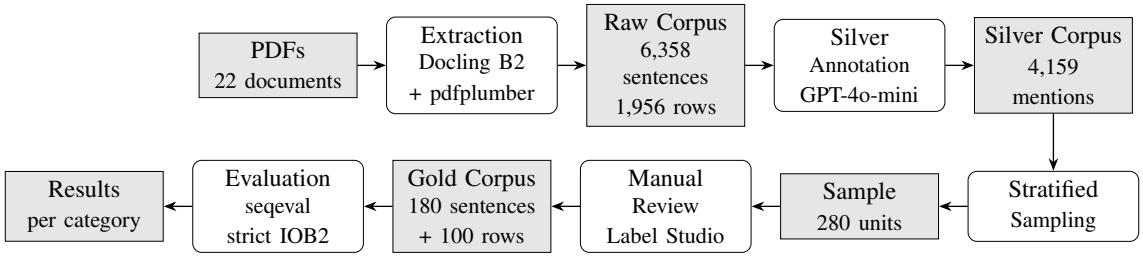


Figure 1. Overview of the EpiCorpus-BR construction pipeline. Grey rectangles indicate data artifacts; white rectangles indicate processing steps.

4.1. Text Extraction from PDF Documents

Extraction applies the **B2 scenario** from [Freitas et al. 2026]: text-only conversion via **Docling** [DS4SD Team 2024] with explicit removal of Markdown tables and preservation of captions. Since the PDFs are natively text-based, OCR is not required. The extracted text passes through three filters: removal of residual tabular lines, conservative sentence segmentation (boundaries detected only before uppercase letters or digits), and a narrative prose filter that discards non-alphabetic fragments and summary lines. The result is **6,358 sentences** in 22 JSONL files (3,343 from SIREVA-SUS and 3,015 from complementary documents).

4.2. Table Row Extraction

The epidemiological tables in SIREVA-SUS concentrate most occurrences of METRICA_EPI and SOROTIPO, categories under-represented in narrative text. A **pdfplumber**-based module extracts these rows in four steps: (1) discarding tables with rotated text (*garbling*); (2) header/data separation by proportion of numeric cells; (3) synthesis of each data row in the format `[caption] | column: value | ...`, making it

self-contained out of context; (4) discarding subtotal and total rows. The module produced **1,956 table rows** in `data/extracted_tables/`, following the same JSONL schema as narrative text with an additional `type: "table"` field.

4.3. Silver Annotation via Few-Shot Prompting

The pipeline adopts **silver annotation** with **GPT-4o-mini** [OpenAI 2024] (temperature 0, output restricted to `json_object`).[1] The system prompt describes the eight categories with a definition and three examples each, plus explicit boundary rules. The few-shot history includes three (sentence, JSON) pairs covering the most frequent patterns: pathogen+serotype, age stratification+laboratory method, and location+temporal period. The model returns spans in the format `{"texto": "...", "tipo": "CATEGORY"}`, which we convert to the IOB2 scheme by character-level alignment. For example, in *"taxa de incidência de meningite por Neisseria meningitidis sorogrupo B"*, the alignment recovers the spans *taxa de incidência* (METRICA_EPI), *meningite* (MANIFESTACAO_CLINICA), *Neisseria meningitidis* (PATOGENO), and *sorogrupo B* (SOROTIPO). Silver annotation covered all 22 documents, producing **4,159 identified entity mentions**.

4.4. Gold Standard Construction

To evaluate the quality of the silver annotation and establish a reliable reference set, the pipeline selects a stratified corpus sample for manual review.

Sampling: narrative text. Sampling is stratified by document, totalling **180 sentences** distributed across the 12 SIREVA-SUS reports. Stratification ensures that all temporal periods and epidemiological patterns of the 2013–2024 historical series are represented in the gold set.

Sampling: table rows. A second gold set was built from silver-annotated table rows, following the same stratified sampling procedure by report, with **100 rows** distributed proportionally across the 12 series years. This set is essential to cover METRICA_EPI, METODO_LAB, and FAIXA_ETARIA in tabular context, categories underrepresented or absent in narrative text.

Manual review. Both sets were exported to **Label Studio** [Tkachenko et al. 2020] in JSON format with silver pre-annotations and reviewed against the tagset guidelines (Section 3). Ambiguous cases were resolved by consulting the guidelines. The reference gold set corresponds to the first annotator's review; where the second annotator's independent review (see the agreement paragraph below) diverged, disagreements were reconciled through discussion in light of the guidelines, and the agreed decision was recorded as the final gold label. For table rows, each numeric value associated with an epidemiological metric was individually annotated as METRICA_EPI (see Section 3.1).

Final format. The text gold set is stored in `data/gold/gold_sample.jsonl` and the table set in

[1] Reproducibility: silver annotation used the `gpt-4o-mini` model; the LLM comparison used `gemini-2.5-flash` (Google AI Studio, accessed July 2026) with the same prompt and temperature 0; the supervised baseline fine-tunes BERTimbau base (110M parameters) for 4 epochs (learning rate 3×10^{-5}, batch 16, max sequence length 192, seed 42).

`data/gold/gold_tables.jsonl`, both following the same field schema as the silver records (`tokens`, `bio_tags`, and `spans`), ensuring direct compatibility with the evaluation module.

Inter-annotator agreement. Guideline consistency was verified by a second domain expert annotator who independently reviewed the same 180 text sentences. Cohen's coefficient [Landis and Koch 1977] over IOB2 labels reached $\kappa = \mathbf{0.76}$ (*substantial agreement*), with 90.5% token-level agreement and a symmetric span F_1 of 0.78. Table 2 details the results per category. PATOGENO achieved the highest κ (0.88), benefiting from the precision of Latin scientific nomenclature. MANIFESTACAO_CLINICA (κ = 0.36), LOCAL (κ = 0.25), and METODO_LAB (κ = 0.21) showed lower values: the second annotator identified, respectively, 48%, 11%, and 20% more mentions in these categories, indicating annotation boundary ambiguity that warrants guideline refinement.

Table 2. Inter-annotator agreement (IAA) per category: Cohen's κ (IOB2 tokens) and symmetric span F_1.

Category	κ	F_1
PATOGENO	0.88	0.94
METRICA_EPI	0.73	0.84
SOROTIPO	0.67	0.57
FAIXA_ETARIA	0.65	0.85
PERIODO_TEMPORAL	0.45	0.71
MANIFESTACAO_CLINICA	0.36	0.69
LOCAL	0.25	0.71
METODO_LAB	0.21	0.55
Overall	**0.76**	**0.78**

Boundary cases and guideline decisions. The three lowest-agreement categories share one source of disagreement: the line between a named entity and a nearby but different concept. For LOCAL (κ = 0.25), the split is place versus institution. The guidelines annotate explicit toponyms and exclude institution names, but acronyms such as *IAL* (Instituto Adolfo Lutz) fall in between when they stand for the geographic node of the surveillance laboratory network. We annotate them as LOCAL only in that locative use, and leave them out when they name the institution as an organization. For METODO_LAB (κ = 0.21), the line is a named technique versus a generic reference: *cultura* and *sorotipagem molecular* are annotated, while a bare *exame* or *laboratório* is not. For MANIFESTACAO_CLINICA (κ = 0.36), the question is how much of a multiword condition to include: we annotate the full named condition (*meningite bacteriana* as a single span) and established domain labels such as *doença invasiva*, but exclude a bare *doença* or *infecção*. These decisions are now recorded as explicit boundary rules in the released guidelines, so that future annotation rounds can apply them from the start.

5. The EpiCorpus-BR

EpiCorpus-BR is built from 12 annual SIREVA-SUS reports (2013–2024) and 10 complementary epidemiological surveillance documents, totalling 22 documents. The SIREVA-SUS sub-corpus totals **3,343 sentences** and **19,762 tokens**, with a mean of 5.9 tokens

per sentence. The low mean reflects short text fragments from captions, lists, and section headings that passed the narrative prose filter, coexisting with full epidemiological sentences. Sentences are spread fairly evenly across the 2013–2024 series, and the 2023 and 2024 reports concentrate the highest volumes (385 and 370 sentences) due to the expansion of surveillance scope during that period.

5.1. Entity Distribution

Of the 3,343 sentences in the SIREVA-SUS silver set, **772 (23.1%)** contain at least one annotated entity, totalling **1,139 mentions** distributed across seven of the eight tagset categories. Table 3 presents the distribution by category in the SIREVA-SUS silver set, alongside the gold-standard counts.

Table 3. Mention distribution per category, in the SIREVA-SUS silver set (3,343 sentences) and in the gold standard (280 units: 180 text sentences + 100 table rows).

	Silver (SIREVA)		Gold (280 units)		
Category	**Ment.**	**%**	**Text**	**Tab.**	**Total**
LOCAL	592	52.0	154	0	154
PATOGENO	379	33.3	54	2	56
FAIXA_ETARIA	57	5.0	46	116	162
METODO_LAB	54	4.7	38	54	92
SOROTIPO	24	2.1	21	0	21
PERIODO_TEMPORAL	18	1.6	5	2	7
MANIFESTACAO_CLINICA	15	1.3	6	31	37
METRICA_EPI	0	0.0	22	638	660
Total	**1,139**	**100.0**	**346**	**843**	**1,189**

LOCAL and PATOGENO account for 85.3% of mentions, consistent with the structure of SIREVA-SUS reports, where surveillance is organized primarily by etiological agent and by federal unit. The absence of METRICA_EPI mentions in narrative text was confirmed as structural: the domain's quantitative metrics (such as strain counts by age group and susceptibility percentages) are concentrated in epidemiological tables, inaccessible to the narrative prose filter. The tabular extraction described in Section 4.2 covers this content; its 100-row gold standard is evaluated in Section 6.

The gold-standard columns of Table 3 confirm the same tabular concentration: METRICA_EPI and FAIXA_ETARIA are predominantly represented in tables, while LOCAL and PATOGENO concentrate in narrative text, reflecting the structure of SIREVA-SUS reports.

5.2. Complementary Documents

Beyond the SIREVA-SUS reports, the corpus incorporates 10 complementary surveillance documents: three *Electronic Epidemiological Bulletins*, five numbered *Epidemiological Bulletins*, and two *Meningitis Advisories*, all from the Ministry of Health or the Health Surveillance Secretariat. They add **3,015 sentences** and **3,020 silver mentions**, and their distinct structure and terminology broaden the linguistic diversity of the corpus; the same extraction and silver-annotation pipeline was applied to them. Their entity

density is much higher than in the SIREVA-SUS sub-corpus, **1.00 mention per sentence** against **0.34**, because the thematic bulletins describe specific cases and outbreaks rather than long-term surveillance prose. The two meningitis advisories are the densest documents, with 56.9% and 48.9% of their sentences carrying at least one entity.

6. Results

The evaluation compares silver annotations against the gold standard using **seqeval** [Nakayama 2018] in strict matching (IOB2) mode. Table 4 presents F_1 per category for the three evaluated configurations: narrative text (180 sentences, 346 mentions), table rows (100 rows, 843 mentions), and combined corpus (280 units, 1,189 mentions).

Table 4. Per-category F_1 (strict seqeval, IOB2): GPT-4o-mini silver annotation in the three configurations, a second-LLM annotator (Gemini 2.5 Flash, same prompt), and the BERTimbau baseline (trained on silver, combined-gold test, Section 6.2). N = support in the combined corpus.

	GPT-4o-mini (silver)			Gemini	BERTimbau	
Category	**Text**	**Tab.**	**Comb.**	**Comb.**	**Comb.**	**N**
PATOGENO	0.97	0.00	0.95	0.96	0.98	56
SOROTIPO	0.92	0.00	0.92	0.62	0.76	21
PERIODO_TEMPORAL	0.89	0.00	0.73	0.62	0.77	7
METODO_LAB	0.88	0.62	0.71	0.65	0.63	92
FAIXA_ETARIA	0.85	0.25	0.43	0.45	0.43	162
LOCAL	0.61	0.00	0.59	0.26	0.30	154
MANIFESTACAO_CLINICA	0.50	0.59	0.58	0.68	0.51	37
METRICA_EPI	0.00	0.91	0.89	0.94	0.89	660
Micro avg	0.73	0.79	0.77	0.77	0.73	1189
Macro avg	0.70	0.34	0.73	0.65	0.66	1189

In narrative text, PATOGENO and SOROTIPO exceed $F_1 = 0.92$, helped by their consistent Latin nomenclature. METRICA_EPI scores zero in text, because its metrics are almost entirely tabular, yet it reaches 0.91 on table rows. FAIXA_ETARIA drops from 0.85 in text to 0.25 in tables, and LOCAL shows only moderate precision (0.57). The combined silver micro-F_1 is 0.77. Section 6.1 analyses the error patterns behind these numbers.

Robustness across LLM annotators. To check whether the annotation quality is specific to GPT-4o-mini, we ran the same prompt with a second, more recent model, Gemini 2.5 Flash, over the same 280 gold units. It reaches the *same* combined micro-F_1 of 0.77 (Table 4). This suggests that the silver annotation is robust to the choice of LLM and is not an artefact of a single model. The two models redistribute the errors rather than removing them. Gemini is stronger on the tabular split (0.86 vs. 0.79), on METRICA_EPI (0.94), and on MANIFESTACAO_CLINICA (0.68), but weaker on SOROTIPO (0.62). Like GPT-4o-mini, it still collapses on the ambiguous LOCAL category (0.26), which shows that this category is a guideline problem, not a model one.

6.1. Error Analysis

Manual inspection of the disagreements between the silver annotation and the gold standard shows that the errors are not spread evenly across categories. They concentrate in

three systematic patterns, and all three are **omissions** (false negatives) rather than confusion between categories. Table 5 gives one representative example of each pattern. Below, we explain what each example is meant to show and how much F_1 it costs (Table 4).

(E1) Omitted numeric values in narrative text. This pattern alone explains the zero F_1 of METRICA_EPI in text. The `(n=166)` example isolates the failure: the model reads the surrounding context correctly, since it annotates *cultura* as METODO_LAB, yet it does not extend the annotation to the adjacent count that the guidelines mark as METRICA_EPI. The problem is therefore recall of the numeric span, not a wrong label, because the same model annotates these counts correctly when they appear alone in table cells ($F_1 = 0.91$).

(E2) Institutional acronyms used as locations. This pattern explains the moderate score of LOCAL (0.59). The *IAL* (Instituto Adolfo Lutz) example shows the boundary the model misses: it captures explicit place names such as *Brasil* and *São Paulo*, but not domain acronyms that stand for a place in context, here the location of the surveillance laboratory network. This is the same LOCAL boundary that drove the low inter-annotator agreement for the category (Section 4.4), which confirms it is a guideline issue, not a model one.

(E3) Compact age formats in tables. This pattern explains why FAIXA_ETARIA falls from $F_1 = 0.85$ in text to 0.25 in tables. The example, an abbreviated age band such as `<1 ano` in a table header, shows that the failure is one of surface form, not of concept: the few-shot examples cover spelled-out ranges (*"menor que 12 meses"*) but not the short notation of tabular headers, so the model omits them or misplaces their boundaries.

Together, the three patterns show that the errors come from domain-specific surface forms that are under-represented in the few-shot prompt, not from conceptual ambiguity between categories. Two routes address this: richer few-shot examples and guidelines, and the supervised model of Section 6.2, which learns these surface forms from the distribution of the silver corpus.

Table 5. The three systematic silver-annotation errors. Each row shows the token that should have been annotated, its gold label, and the silver output. O means the model left the token unannotated, so all three patterns are omissions (false negatives).

Type	Token (in context)	Gold label	Silver
E1: omitted value	`(n=166)` in "…por cultura `(n=166)`"	B-METRICA_EPI	O
E2: locative acronym	*IAL* in "o *IAL* monitora … os estados"	B-LOCAL	O
E3: compact age	`<1 ano` in a tabular age-group header	B-FAIXA_ETARIA	O

6.2. A Supervised Baseline under Silver Supervision

To assess whether the corpus can train a dedicated Portuguese NER model, and not only measure the LLM annotator, we fine-tune BERTimbau [Souza et al. 2020] (base, cased)

as a token classifier under *silver (distant) supervision*. The model is trained on the silver-annotated corpus (narrative text and table rows) with the 280 gold units removed to prevent test leakage (8,034 training units), and evaluated on the gold standard using the same seqeval strict-IOB2 protocol as in Section 6. Word-level labels are aligned to the first subword token; we train for 4 epochs with learning rate 3×10^{-5}, batch size 16, maximum sequence length 192, and a fixed seed for reproducibility. This setup directly tests the downstream utility of the corpus while keeping the gold set untouched during training.

The rightmost column of Table 4 reports the results. Trained only on the silver corpus, BERTimbau reaches a combined micro-F_1 of **0.73**, close to the 0.77 of the silver annotation it learned from. This shows that the corpus is usable to train a dedicated model and that its silver labels are consistent enough to be learned. The model is strongest where the annotation is terminologically regular, namely PATOGENO (0.98, Latin nomenclature) and METRICA_EPI (0.89, and 0.90 on table rows alone). It also matches the silver micro-F_1 on the tabular split (0.79), which confirms that the tabular sub-corpus transfers well to a trained model. Performance drops on the ambiguous categories, with LOCAL falling to 0.30. This mirrors error type E2 (Section 6.1) and confirms that the bottleneck is institutional and locative ambiguity, not model capacity. Because the model is supervised by the silver labels themselves, it cannot exceed the silver ceiling by much. Reaching 0.73 without ever seeing the gold is therefore a lower bound, and a larger, gold-supervised training set should raise it.

7. Limitations

The gold standard covers only 3.4% of the corpus (280 of 8,314 units); although stratified, the evaluation reflects this subset, and the silver annotation comes from a single model (GPT-4o-mini). The IAA (κ = 0.76, Section 4.4) used two annotators who both started from the silver pre-annotations, which may inflate it relative to annotation from scratch, and the low κ for LOCAL, METODO_LAB, and MANIFESTACAO_CLINICA reflects residual boundary ambiguity. More importantly, the review corrects and reclassifies the model's spans but does not systematically add entities the model never proposed, so the gold standard measures the *precision* of the silver annotation reliably while its true *recall* may be overestimated.

8. Conclusion

This paper presented EpiCorpus-BR, to the best of our knowledge the first public NER corpus for epidemiological surveillance documents in Portuguese. It brings together 8,314 textual units from 22 SIREVA-SUS and Ministry of Health documents, a tagset of eight specialized categories absent from existing Portuguese resources [Oliveira et al. 2022], and a gold standard of 280 manually reviewed units. The combined evaluation (micro-F_1 = 0.77) showed that METRICA_EPI, at zero F_1 in narrative text, reaches 0.89 in tables, so the tabular sub-corpus is essential for a representative evaluation of the domain. A BERTimbau baseline trained on the silver corpus, with the gold held out, reached micro-F_1 = 0.73, confirming that the corpus supports training a dedicated NER model.

As future work, we plan to integrate the three annotation layers already built for SIREVA-SUS documents, Universal Dependencies [Freitas et al. 2026], named entities (this work), and RAG-based semantic retrieval [Freitas et al. 2025], into a unified analysis pipeline.

Availability

EpiCorpus-BR and the pipeline code are publicly available at `https://github.com/ChristianSF/EpiCorpus-BR`, in JSONL format (fields `tokens`, `bio_tags` in IOB2, and `spans`) with documentation and the extraction, annotation, and evaluation modules. The corpus is released under CC BY 4.0 and the code under MIT.

Acknowledgements

The authors thank Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for financial support.

Ethics Statement

The documents used are official publications made available by the Oswaldo Cruz Foundation, the Ministry of Health, and the Health Surveillance Secretariat without access restrictions. The corpus contains no patient personal data, consisting exclusively of epidemiological text aggregated at the population level. During the preparation of this article, generative artificial intelligence tools were used to support language and structure review; the authors reviewed and are responsible for the final content of the publication.

References

Albuquerque, H. O., Souza, E., Gomes, C., Pinto, M. H. d. C., Filho, R. P. S., Costa, R., Lopes, V. T. d. M., da Silva, N. F. F., de Carvalho, A. C., and Sampaio, A. d. O. (2023). Named entity recognition: a survey for the Portuguese language. *Procesamiento del Lenguaje Natural*, 70:171–185.

Brito, M., Pinheiro, V., Furtado, V., Monteiro Neto, J. A., Bomfim, F. d. C. J., da Costa, A. C. F., and Silveira, R. (2023). CDJUR-BR – uma coleção dourada do judiciário brasileiro com entidades nomeadas refinadas. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, pages 177–186, Porto Alegre. Sociedade Brasileira de Computação.

Ding, B., Qin, C., Zhao, R., Luo, T., Li, X., Chen, G., Xia, W., Hu, J., Luu, A. T., and Joty, S. (2024). Data augmentation using LLMs: Data perspectives, learning paradigms and challenges. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 8764–8786.

Doğan, R. I., Leaman, R., and Lu, Z. (2014). NCBI disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10.

DS4SD Team (2024). Docling technical report. *arXiv preprint arXiv:2408.09869*.

Freitas, C., Rabonato, R. T., and Berton, L. (2025). Enhancing epidemiological insights with RAG for SIREVA-SUS reports. In *Anais do XXII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, pages 1364–1375, Fortaleza/CE, Brazil. Sociedade Brasileira de Computação.

Freitas, C., Real, L., Berton, L., and Paiva, V. d. (2026). Towards a Universal Dependencies corpus for Portuguese epidemiological reports. In *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 2*, pages 228–237, Salvador, Brazil. Association for Computational Linguistics.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Moreira, H., Silva, P. F. d., Vieira, R., and Moreira, V. (2025). PetroGeoNER: um conjunto de dados refinado e unificado para NER no domínio de petróleo e gás. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, pages 259–271, Porto Alegre. Sociedade Brasileira de Computação.

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Nakayama, H. (2018). seqeval: A python framework for sequence labeling evaluation. `https://github.com/chakki-works/seqeval`.

Oliveira, L. E. S. e., Peters, A. C., da Silva, A. M. P., Gebeluca, C. P., Gumiel, Y. B., Cintho, L. M. M., Carvalho, D. R., Al Hasan, S., and Moro, C. M. C. (2022). SemClinBr – a multi-institutional and multi-specialty semantically annotated corpus for Portuguese clinical NLP tasks. *Journal of Biomedical Semantics*, 13(1).

OpenAI (2024). GPT-4o mini: Advancing cost-efficient intelligence. `https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/`.

Piai, L., Di-Felippo, A., and Romano, N. T. (2025). Entidades nomeadas em tweets do mercado de ações: uma anotação refinada e com motivação linguística. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, pages 654–663, Porto Alegre. Sociedade Brasileira de Computação.

Pontes, M. F., Pedrosa, R. C., Lopes, P. H., and Luz, E. J. (2024). Avaliando aprendizagem federada com criptografia homomórfica para reconhecimento de entidades médicas nomeadas utilizando modelos compactos de BERT. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, pages 48–56, Porto Alegre. Sociedade Brasileira de Computação.

Ramos, F. d. O., Marvila, C. G., and Colombo, C. d. S. (2026). Uma abordagem híbrida para reconhecimento de entidades nomeadas em relatos de casos clínicos. In *Anais do XXII Simpósio Brasileiro de Sistemas de Informação (SBSI)*, pages 216–233, Porto Alegre. Sociedade Brasileira de Computação.

Schneider, E. T. R., de Souza, J. V. A., Knafou, J., Oliveira, L. E. S. e., Copara, J., Gumiel, Y. B., Oliveira, L. F. A. d., Paraiso, E. C., Teodoro, D., and Barra, C. M. C. M. (2020). BioBERTpt – a Portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 65–72. Association for Computational Linguistics.

Secretaria de Vigilância em Saúde (2024). Vigilância epidemiológica. Ministério da Saúde do Brasil. Disponível em: `https://www.gov.br/saude/pt-br/composicao/svs`. Acesso em: mai. 2026.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: Pretrained BERT models for Brazilian Portuguese. In *Intelligent Systems (BRACIS 2020)*, pages 403–417. Springer.

Tkachenko, M., Malyuk, M., Holmberg, N., Liubimov, N., and Strulov, A. (2020). Label Studio: Data labeling software. `https://github.com/heartexlabs/label-studio`.

Uzuner, Ö., South, B. R., Shen, S., and DuVall, S. L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Curupira: Arquiteturas Híbridas Transformer-BiLSTM para Detecção de Discurso Ofensivo em Português Brasileiro

Gabrielly Alves Gomes[1], **Iago de Sousa Aragão**[1]
Patricia Medyna Lauritzen de Lucena Drumond[1]

[1]Universidade Federal do Piauí (UFPI)
Campus Universitário Ministro Petrônio Portella - Ininga, Teresina - PI, 64049-550, Brasil

{gabrielly.gomes,iago.aragao,patriciamedyna}@ufpi.edu.br

Abstract. *This work presents Curupira, a controlled empirical study of neural architectures for offensive speech detection in Brazilian Portuguese under a rigorous methodological protocol: honest train/validation/test split, five random seeds, explicit parameter-capacity control, and paired bootstrap significance testing. Eight configurations are evaluated on OLID-BR, spanning four orders of magnitude in parameters. The results should be interpreted for this dataset and training regime, and indicate five main findings: (i) performance clusters around Macro-F1 close to 0.70, with BERTimbau (110M parameters) surpassing a lightweight Transformer→BiLSTM (350K) by 4.3 points; (ii) global attention provides statistically significant gains over a capacity-matched BiLSTM; (iii) increasing neural capacity does not improve performance in the tested setting; (iv) composition order matters, with Transformer→BiLSTM surpassing BiLSTM→Transformer by 6.3 points; (v) sinusoidal positional encoding over frozen embeddings substantially degrades performance, reducing Macro-F1 from 0.66 to 0.26.*

Resumo. *Este trabalho apresenta o Curupira, estudo empírico controlado de arquiteturas neurais para detecção de discurso ofensivo em português brasileiro sob protocolo metodológico rigoroso:* split *treino/validação/teste honesto, cinco sementes aleatórias, controle explícito de capacidade paramétrica e teste de significância via* bootstrap *pareado. Oito configurações são avaliadas no OLID-BR, cobrindo quatro ordens de magnitude em parâmetros. Os resultados devem ser interpretados para este conjunto de dados e regime de treinamento, e indicam cinco achados principais: (i) o desempenho se concentra próximo de Macro-F1 $=$ 0,70, com o BERTimbau (110M de parâmetros) superando um híbrido Transformer→BiLSTM leve (350K) em 4,3 pontos; (ii) a atenção global produz ganho estatisticamente significativo sobre uma BiLSTM equiparada em capacidade; (iii) aumentar a capacidade neural não melhora o desempenho no cenário avaliado; (iv) a ordem de composição importa, com Transformer→BiLSTM superando BiLSTM→Transformer em 6,3 pontos; (v) codificação posicional senoidal sobre* embeddings *congelados degrada substancialmente o desempenho, reduzindo o Macro-F1 de 0,66 para 0,26.*

1. Introdução

A proliferação de conteúdo ofensivo em plataformas digitais constitui problema relevante para comunidades *online* e para a sociedade, o que torna a moderação automatizada baseada em aprendizado de máquina uma estratégia necessária diante do volume de

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 163–174
October 19-22, 2026 ©2026 Association for Computational Linguistics

publicações em redes sociais. Estudos sobre linguagem ofensiva e discurso de ódio em redes sociais, incluindo Twitter/X, mostram tanto a viabilidade quanto a complexidade da tarefa: distinguir *hate speech* de linguagem meramente ofensiva exige sensibilidade contextual, especialmente diante de coloquialidade, ironia e diferentes alvos de ataque [2, 19]. Embora a pesquisa tenha avançado para a língua inglesa, com recursos como o OLID [22] e as *shared tasks* OffensEval, o cenário para o português brasileiro permanece comparativamente menos explorado, conforme documentam estudos sobre ToLD-Br [12], HateBR [20] e anotação hierárquica de discurso de ódio em português [8].

Três lacunas convergentes motivam este trabalho. A primeira é a avaliação de arquiteturas híbridas Transformer-BiLSTM sob condições de baixo recurso, com *embeddings* estáticos, conjuntos de dados pequenos e ausência de pré-treino, condições frequentes em aplicações reais para PT-BR. A segunda é que trabalhos anteriores nem sempre controlam capacidade paramétrica, o que pode confundir efeito arquitetural com efeito de escala, problema sinalizado por Ramos et al. [16]. A terceira é que práticas comuns como uso de semente única e ausência de testes pareados comprometem a robustez das conclusões; Reimers et al. [17] demonstraram que a variância entre sementes pode inverter conclusões, e Dror et al. [5] argumentam que testes pareados devem constituir padrão mínimo.

Este trabalho apresenta o projeto Curupira[1], cujas contribuições podem ser sintetizadas em cinco eixos. A primeira contribuição (C1) é caracterizar, no OLID-BR e sob o protocolo avaliado, um agrupamento de desempenho próximo de Macro-F1 $= 0{,}70$ em oito modelos que cobrem quatro ordens de magnitude em parâmetros. A segunda (C2) é mostrar que o híbrido Transformer$\rightarrow$BiLSTM supera uma BiLSTM equiparada em 350K parâmetros com significância estatística via *bootstrap* pareado. A terceira (C3) é documentar que aumentar a capacidade dos modelos neurais leves não produziu ganho significativo neste cenário. A quarta (C4) é uma ablação de codificação posicional que mostra forte degradação quando PE senoidal é somada a *embeddings* congelados. A quinta (C5) é formalizar um protocolo metodológico replicável, com separação validação/teste, múltiplas sementes e comparação pareada.

2. Trabalhos Relacionados

2.1. Detecção de Discurso Ofensivo e Desbalanceamento

A detecção automática de discurso ofensivo evoluiu de métodos baseados em *features* manuais, como TF-IDF e n-gramas combinados com SVM ou regressão logística, para modelos de aprendizado profundo, conforme sintetizado por Fortuna e Nunes [7]. Estudos recentes indicam o predomínio de modelos baseados em *Transformers*, embora CNNs, LSTMs e arquiteturas híbridas ainda permaneçam competitivos em cenários com conjuntos de dados menores [13, 16]. Outro aspecto central é o desbalanceamento de classes: quando não tratado, a acurácia global tende a mascarar o desempenho da classe minoritária, o que justifica a adoção de Macro-F1 como métrica principal [12, 22].

2.2. BiLSTM, Transformer e Arquiteturas Híbridas

BiLSTMs e *Transformers* representam duas estratégias complementares para modelagem textual: a primeira enfatiza dependências sequenciais bidirecionais, enquanto a segunda

[1] *Curupira*: figura do folclore brasileiro, protetor das florestas, cujos pés voltados para trás confundem invasores. O nome evoca a função do modelo, reconhecer ataques e proteger o espaço digital.

utiliza *self-attention* para capturar relações globais entre *tokens* [9, 21]. Trabalhos anteriores exploraram a combinação dessas famílias em arquiteturas híbridas. Huang et al. [10], com o TRANS-BLSTM, e Khan et al. [11], com o BiCHAT, reportam ganhos em tarefas de entendimento de linguagem e detecção de discurso de ódio. Ainda assim, esses resultados costumam ser obtidos em condições mais favoráveis, com pré-treino massivo, *embeddings* treináveis ou maior escala de dados, e frequentemente sem controle explícito de capacidade paramétrica [16].

2.3. Português Brasileiro e Rigor Experimental

Os recursos para detecção de discurso ofensivo em português incluem o OLID-BR, adaptação do esquema OLID [23], o ToLD-Br apresentado por Leite et al. [12] com aproximadamente 21 mil *tweets*, o HateBR disponibilizado por Vargas et al. [20], com anotação especializada sobre comentários do Instagram, e o conjunto hierárquico de Fortuna et al. [8]. Estudos recentes também analisam repositórios de toxicidade em português [14] e o uso de LLMs na detecção de discurso de ódio em português [15]. O trabalho de Leite et al. [12] é particularmente relevante ao estabelecer que *transfer learning* do inglês não supera modelos monolíngues, que *zero-shot learning* alcança apenas Macro-F1 de 0,56 e que pelo menos seis mil exemplos são necessários para resultados confiáveis em PT-BR. O BERTimbau, disponibilizado por Souza et al. [18], estabeleceu-se como referência para a língua, com Macro-F1 de 0,76 no ToLD-Br em classificação binária.

Quanto ao rigor experimental, Reimers et al. [17] documentaram que resultados reportados com única semente não constituem evidência confiável, visto que a variância entre inicializações pode ser maior do que as diferenças supostas entre modelos. Baseados nessa evidência, Dror et al. [5] argumentam que a adoção de testes pareados deve ser padrão em NLP, e recomendam explicitamente *bootstrap* pareado para situações com poucas observações, nas quais o teste de Wilcoxon *signed-rank* sofre de baixa potência estatística. O *framework* sistemático para comparação de múltiplos classificadores oferecido por Demšar [3] e os métodos de *bootstrap* não paramétrico formalizados por Efron e Tibshirani [6] fornecem a base teórica do protocolo adotado na Seção 3.

3. Metodologia

3.1. Conjunto de Dados OLID-BR

O OLID-BR contém 6.472 *tweets* em português brasileiro, anotados para detecção binária de conteúdo ofensivo (OFF versus NOT). A análise exploratória conduzida na fase preliminar deste trabalho identificou 354 amostras rotuladas como OFF, porém sem *toxic_spans* válidos. Os *toxic_spans* são marcações de trechos específicos do texto identificados pelos anotadores como ofensivos, exigidas pelo esquema original do OLID-BR para todos os exemplos rotulados como OFF; sua ausência em amostras OFF configura inconsistência formal de anotação. Essas 354 amostras foram removidas do conjunto de treino, com base exclusivamente em propriedades dos metadados e antes de qualquer experimento arquitetural, o que exclui *data leakage* retroativo. O conjunto de teste foi preservado integralmente, de modo a refletir a distribuição original. Um experimento de sensibilidade conduzido sobre os três modelos centrais (BiLSTM-64, BiLSTM-128 e T$\rightarrow$B-*small*, cinco sementes cada) mostrou que a filtragem produz diferenças de Macro-F1 médio inferiores a $0{,}02$ pontos, sem alterar o ranking arquitetural ou as conclusões comparativas reportadas; a contribuição mais relevante da filtragem é a redução

da variância entre sementes, particularmente para o BiLSTM-128 (σ cai de $0{,}076$ para $0{,}032$). A Tabela 1 resume a distribuição resultante.

A diferença marginal de $-0{,}002$ pontos observada para o T$\rightarrow$B-*small* situa-se dentro do erro de medição entre execuções, e o ranking arquitetural mantém-se inalterado em ambos os cenários.

O pré-processamento aplicado segue convenção comum na literatura para texto de redes sociais: conversão para minúsculas, remoção de URLs, menções de usuário, emojis, caracteres não alfabéticos (com preservação de diacríticos portugueses) e padrões de risadas frequentes (por exemplo, sequências como `kkk`, `haha` e `rsrs`). A tokenização adota vocabulário de 50 mil palavras e comprimento máximo de 100 *tokens*.

Tabela 1. Distribuição do OLID-BR antes e após filtragem de amostras OFF sem *toxic_spans* válidos.

	Sem filtragem			Com filtragem		
Conjunto	Total	OFF (%)	NOT (%)	Total	OFF (%)	NOT (%)
Treino	4.860	84,3	15,7	4.506	83,1	16,9
Validação (10%)	486	84,3	15,7	451	83,1	16,9
Teste	1.732	85,5	14,5	1.732	85,5	14,5

3.2. Protocolo de *Split* e Seleção de Modelo

Uma decisão metodológica central deste trabalho é a separação estrita entre conjuntos de validação e teste. Dez por cento do conjunto de treino foram reservados como conjunto de validação, estratificado por rótulo com semente fixa para garantir reprodutibilidade da partição. Todos os *callbacks* de seleção de modelo (`EarlyStopping`, `ModelCheckpoint` e `ReduceLROnPlateau`) operam exclusivamente sobre a perda de validação; o conjunto de teste é tocado uma única vez por execução, ao final do treinamento, apenas para reportar a métrica final. Essa disciplina evita a contaminação sutil, porém significativa, decorrente de selecionar a época ótima com base em desempenho no teste, conforme discutido por Dror et al. [5].

3.3. *Embeddings* e Ponderação de Classes

Os *embeddings* FastText pré-treinados em CommonCrawl para o português, disponibilizados por Bojanowski et al. (2017) com $d = 300$, são utilizados como camada de entrada dos modelos neurais leves e mantidos congelados (`trainable=False`) durante todo o treinamento. Essa decisão isola o efeito da composição arquitetural, impedindo que ganhos aparentes derivem de adaptação dos próprios *embeddings*. O desbalanceamento de classes é tratado por ponderação inversa à frequência, seguindo a fórmula $w_c = N/(k \cdot n_c)$, onde N é o total de amostras, k o número de classes e n_c a contagem da classe c. Sobre o treino filtrado, os pesos resultantes são $w_{\text{NOT}} = 2{,}96$ e $w_{\text{OFF}} = 0{,}60$.

3.4. Arquiteturas Avaliadas

Oito configurações foram avaliadas, organizadas em cinco famílias.

Baselines TF-IDF. Dois classificadores lineares foram treinados sobre vetores TF-IDF com n-gramas de tamanho um e dois, limitados a cinquenta mil *features*, o

primeiro baseado em regressão logística e o segundo em SVM linear, ambos com `class_weight='balanced'`. A função dessas referências é fornecer piso informativo para a comparação arquitetural posterior, respondendo à recomendação de Fortuna e Nunes [7] sobre o valor residual de *baselines* simples em regimes de baixo recurso.

BiLSTM puro. Duas configurações foram avaliadas, com sessenta e quatro (*BiLSTM-64*) e cento e vinte e oito unidades (*BiLSTM-128*). Cada variante recebe a sequência de *embeddings* FastText congelados, aplica uma camada BiLSTM com `return_sequences=True`, seguida de `GlobalMaxPool1D`, `Dense`(32, ReLU), `Dropout`(0,5) e `Dense`(1, sigmoide). As variantes totalizam aproximadamente 200K e 350K parâmetros, respectivamente, isolando o efeito da capacidade recorrente pura.

Híbrido Transformer→BiLSTM (T→B). A variante inspirada no TRANS-BLSTM [10] posiciona um bloco atencional antes da camada recorrente, na seguinte ordem: *embedding* congelado, Pre-LayerNorm, *MultiHeadAttention* com quatro cabeças e $d_k = 64$ com conexão residual, BiLSTM com `return_sequences=True` e `GlobalMaxPool1D`. Duas versões compartilham essa estrutura, diferenciando-se apenas no tamanho da camada recorrente: a *small* (BiLSTM de 64 unidades, aproximadamente 350K parâmetros) e a *large* (128 unidades, aproximadamente 500K). Nenhuma das duas inclui codificação posicional por padrão; o impacto dessa decisão é avaliado como ablação específica na Seção 4.

Híbrido BiLSTM→Transformer (B→T). A variante com ordem invertida aplica primeiro a BiLSTM (sessenta e quatro unidades, `return_sequences=True`) e em seguida o mesmo bloco atencional, totalizando aproximadamente 350K parâmetros. A comparação direta com a T→B-*small*, equiparada em capacidade paramétrica, isola o efeito da ordem de composição.

BERTimbau. O modelo `neuralmind/bert-base-portuguese-cased`, disponibilizado por Souza et al. [18], foi submetido a *fine-tuning* completo durante três épocas, com *batch* de oito, taxa de aprendizagem de 2×10^{-5}, precisão mista (`fp16`) e comprimento máximo de 128 *tokens*. Para manter consistência de protocolo com os modelos Keras, implementou-se um `WeightedLossTrainer` customizado sobre o `Trainer` do HuggingFace, aplicando *cross-entropy* ponderada pelos mesmos pesos de classe definidos para os modelos leves.

3.5. Configuração Experimental e Análise Estatística

Todos os modelos neurais foram treinados com otimizador Adam, perda *binary cross-entropy*, *batch size* 32 e máximo de dez épocas, com cinco sementes aleatórias por configuração no conjunto $\{42, 123, 456, 789, 1024\}$. Os experimentos foram executados em uma GPU NVIDIA A100 através do Google Colab, utilizando TensorFlow/Keras para os modelos leves e HuggingFace *Transformers* para o BERTimbau.

Para avaliação de significância estatística entre pares de modelos, adotou-se *bootstrap* pareado não paramétrico, conforme recomendado por Dror et al. [5] para cenários com número limitado de observações. Dadas as cinco medições pareadas por semente, o vetor de diferenças foi reamostrado com reposição 10 mil vezes, computando-se o intervalo de confiança de 95% da média da diferença. Uma diferença foi considerada estatisticamente significativa quando o intervalo de confiança não contém zero. A escolha do *bootstrap* em vez do Wilcoxon *signed-rank* se justifica pela observação de que o valor-p

mínimo alcançável pelo Wilcoxon com $n = 5$ é $0{,}0625$, tornando o teste incapaz de rejeitar a hipótese nula mesmo na presença de diferenças reais, problema discutido por Dror et al. [5] e compatível com as recomendações de Efron e Tibshirani [6] para inferência em amostras pequenas. O desvio-padrão σ de Macro-F1 entre sementes é reportado com intervalo de confiança de 95% obtido por *bootstrap* sobre a amostra das cinco sementes.

4. Resultados

4.1. Ranking Final e Significância Estatística

A Tabela 2 consolida o desempenho das oito configurações no conjunto de teste, ordenadas por Macro-F1 médio.

Tabela 2. Ranking final no conjunto de teste (média $\pm\,\sigma$ sobre 5 sementes).

#	Modelo	Macro-F1	NOT-F1	OFF-F1	Parâm.
1	BERTimbau	$\mathbf{0{,}705 \pm 0{,}012}$	$\mathbf{0{,}508 \pm 0{,}008}$	$\mathbf{0{,}901 \pm 0{,}022}$	$\sim$110M
2	Híbrido T$\rightarrow$B-*small*	$0{,}662 \pm 0{,}008$	$0{,}478 \pm 0{,}007$	$0{,}844 \pm 0{,}014$	$\sim$350K
3	Híbrido T$\rightarrow$B-*large*	$0{,}656 \pm 0{,}013$	$0{,}469 \pm 0{,}016$	$0{,}834 \pm 0{,}028$	$\sim$500K
4	BiLSTM-64	$0{,}630 \pm 0{,}017$	$0{,}440 \pm 0{,}011$	$0{,}820 \pm 0{,}023$	$\sim$200K
5	TF-IDF + LR	$0{,}619$	$0{,}365$	$0{,}873$	$\sim$50K
6	BiLSTM-128	$0{,}613 \pm 0{,}032$	$0{,}422 \pm 0{,}031$	$0{,}804 \pm 0{,}039$	$\sim$350K
7	TF-IDF + SVM	$0{,}600$	$0{,}297$	$0{,}903$	$\sim$50K
8	Híbrido B$\rightarrow$T-*small*	$0{,}598 \pm 0{,}044$	$0{,}405 \pm 0{,}052$	$0{,}791 \pm 0{,}031$	$\sim$350K

Vale destacar um padrão interessante nas métricas por classe: o TF-IDF+SVM apresenta o maior OFF-F1 de todos os modelos avaliados (0,903), superando inclusive o BERTimbau (0,901). O ganho vem, contudo, ao custo do menor NOT-F1 do conjunto (0,297), indicando estratégia de classificação enviesada para a classe majoritária. Essa observação valida empiricamente a adoção de Macro-F1 como métrica principal: otimização da classe majoritária isoladamente pode produzir modelos enganosamente competitivos em métricas agregadas como acurácia, mascarando desempenho insatisfatório na classe minoritária.

O BERTimbau lidera o ranking com $0{,}705 \pm 0{,}012$, seguido pelo híbrido T$\rightarrow$B-*small* com $0{,}662 \pm 0{,}008$. A configuração B$\rightarrow$T-*small* ocupa a última posição entre os modelos neurais, com $0{,}598 \pm 0{,}044$, abaixo inclusive dos *baselines* TF-IDF. A amplitude total entre o melhor e o pior modelo é de aproximadamente 10,7 pontos percentuais, apesar da cobertura de quatro ordens de magnitude em número de parâmetros, de 50 mil (TF-IDF) a 110 milhões (BERTimbau).

A Tabela 3 apresenta os testes de significância via *bootstrap* pareado, conforme o protocolo recomendado por Dror et al. [5]. Os resultados são discutidos em detalhe na Seção 5.

4.2. Métricas por Classe

A classe minoritária NOT constitui gargalo universal. Mesmo o BERTimbau atinge apenas NOT-F1 $= 0{,}508$, e todos os modelos exibem NOT-F1 entre 0,30 e 0,51. O

Tabela 3. *Bootstrap* pareado das diferenças de Macro-F1 (A − B), 10 mil reamostragens.

A	B	Diferença	IC 95%	Sig. ($\alpha = 0{,}05$)
BERTimbau	T→B-*small*	$+0{,}043$	$[+0{,}033;\ +0{,}053]$	✓
BERTimbau	BiLSTM-128	$+0{,}092$	$[+0{,}071;\ +0{,}114]$	✓
T→B-*small*	BiLSTM-128	$+0{,}049$	$[+0{,}027;\ +0{,}071]$	✓
T→B-*small*	B→T-*small*	$+0{,}063$	$[+0{,}030;\ +0{,}097]$	✓
T→B-*large*	BiLSTM-128	$+0{,}040$	$[+0{,}015;\ +0{,}062]$	✓
T→B-*small*	T→B-*large*	$+0{,}009$	$[-0{,}002;\ +0{,}020]$	n.s.
BiLSTM-128	BiLSTM-64	$-0{,}017$	$[-0{,}048;\ +0{,}007]$	n.s.

T→B-*small* obtém NOT-F1 de $0{,}478$, superior ao BiLSTM-128 ($0{,}422$) e ao BiLSTM-64 ($0{,}440$). O OFF-F1 é uniformemente alto entre os modelos neurais (acima de 0,79), refletindo o predomínio da classe majoritária no conjunto de teste.

4.3. Ablação: Codificação Posicional e *Layer Normalization*

A Tabela 4 resume a ablação de codificação posicional. A adição de PE senoidal reduz o Macro-F1 de $0{,}657 \pm 0{,}009$ para $0{,}260 \pm 0{,}183$, com colapso completo da classe minoritária em três de cinco sementes (NOT-F1 igual a zero). Trata-se do efeito arquitetural mais expressivo observado no trabalho, tanto em magnitude quanto em consistência da degradação.

Uma ablação complementar compara Pre-LN, adotada como padrão, e Post-LN sobre a mesma variante T→B-*small*. Os resultados são virtualmente indistinguíveis ($0{,}662 \pm 0{,}008$ contra $0{,}664 \pm 0{,}005$), com intervalo de confiança da diferença pareada $[-0{,}009;\ +0{,}004]$, indicando que a escolha de *LayerNorm* não altera o desempenho neste regime.

Tabela 4. Ablação de *positional encoding* sobre o T→B-*small* (5 sementes).

Configuração	Macro-F1	NOT-F1	OFF-F1	Acc.
Sem PE	$0{,}657 \pm 0{,}009$	$0{,}474 \pm 0{,}008$	$0{,}840 \pm 0{,}011$	$0{,}755 \pm 0{,}014$
Com PE senoidal	$0{,}260 \pm 0{,}183$	$0{,}101 \pm 0{,}140$	$0{,}420 \pm 0{,}506$	$0{,}429 \pm 0{,}395$

5. Discussão

5.1. Desempenho no OLID-BR e Custo-Benefício dos Modelos Leves

Os oito modelos avaliados, cobrindo quatro ordens de magnitude em parâmetros, ficaram em um intervalo relativamente estreito de Macro-F1, entre 0,60 e 0,71. O BERTimbau, com aproximadamente 110 milhões de parâmetros, supera o T→B-*small* ($\sim$350 mil parâmetros) em 4,3 pontos percentuais. Esses resultados não demonstram um limite geral do OLID-BR, mas sugerem que, sob o protocolo experimental adotado e com cerca de 4,5 mil amostras de treino após a filtragem descrita na Seção 3, os ganhos por aumento de capacidade ou refinamento arquitetural são limitados. Essa leitura é compatível com Leite et al. [12], que identificaram seis mil exemplos como piso para resultados confiáveis

em PT-BR. Como implicação prática, a agregação entre OLID-BR, ToLD-Br e HateBR [12, 20] permanece uma direção importante para avaliar se o padrão observado se mantém em corpora maiores e mais diversos.

Esse padrão também tem implicação prática de custo-benefício. O classificador TF-IDF seguido de regressão logística atinge Macro-F1 de 0,619 em menos de dez segundos de treinamento, sem uso de GPU, o que corresponde a aproximadamente 87,8% do desempenho do BERTimbau utilizando cerca de 0,05% dos parâmetros. Em cenários com restrição computacional severa, incluindo moderação em tempo real, sistemas *edge* e *pipelines* de alto volume, a diferença de 8,6 pontos percentuais pode ser aceitável quando se consideram o custo marginal por inferência e os requisitos de latência.

5.2. A Atenção Importa

O achado mais relevante da Tabela 3 é a superioridade estatisticamente significativa do T→B-*small* sobre o BiLSTM-128, dado que ambos possuem a mesma ordem de parâmetros (∼350K). Em experimentos sem controle explícito de capacidade, seria natural atribuir o ganho dos híbridos ao número maior de parâmetros, interpretação contra a qual Ramos et al. [16] alertam em sua revisão ao discutir *ensembles* BiLSTM-*Transformer*. Os resultados aqui apresentados indicam que, neste regime experimental, o mecanismo de *self-attention* contribui para o desempenho ao modelar coocorrências globais entre *tokens* que a recorrência bidirecional local pode não capturar. Esse resultado dialoga, em escala de parâmetros muito menor, com a observação qualitativa de Khan et al. [11] sobre os ganhos da atenção hierárquica no BiCHAT. Ainda assim, trata-se de evidência condicionada ao OLID-BR, aos *embeddings* congelados e às arquiteturas avaliadas, não de uma conclusão universal sobre atenção global em classificação textual.

5.3. Saturação Paramétrica e Variância entre Sementes

Dobrar a capacidade do BiLSTM (de 64 para 128 unidades) não produziu ganho neste protocolo e, inclusive, associa-se a leve deterioração do desempenho médio (diferença de $-0{,}017$, IC 95% $[-0{,}048;\ +0{,}007]$). De modo análogo, o T→B-*large* (∼500K parâmetros) não supera o T→B-*small* (∼350K). Dois mecanismos concorrentes podem explicar esse padrão. O primeiro é saturação informacional local: em regime com poucos milhares de exemplos, a capacidade mínima necessária pode já ser satisfeita pelos modelos menores, de modo que parâmetros adicionais não recebem sinal útil suficiente. O segundo é *overfitting* ao ruído do conjunto de dados: capacidade adicional pode permitir ao modelo memorizar padrões espúrios do treino, produzindo deterioração líquida no teste, apesar da manutenção de *dropout*, *early stopping* e ponderação de classes. A combinação desses dois efeitos é uma hipótese plausível para a direção observada, mas sua confirmação exige avaliação em outros conjuntos de dados e regimes de treinamento.

Vale destacar que o BiLSTM-128 varia de 0,571 a 0,652 entre sementes, amplitude de oito pontos percentuais. Um relato com a semente 1024 concluiria que o BiLSTM-128 é competitivo; com a semente 456, concluiria que é deficiente. Essa vulnerabilidade, já documentada por Reimers et al. [17] e sistematizada por Dror et al. [5], reforça a necessidade de reportar resultados agregados sobre múltiplas sementes como prática metodológica padrão.

5.4. A Ordem de Composição Importa

O T→B-*small* supera o B→T-*small* em 6,3 pontos percentuais (IC 95% $[+0{,}030;\ +0{,}097]$) e apresenta variância substancialmente menor. Os resultados sugerem que, neste regime com *embeddings* FastText congelados, aplicar atenção antes da recorrência é mais adequado, possivelmente porque o bloco atencional opera sobre representações mais estáveis antes da variabilidade introduzida pela BiLSTM. Assim, além do tipo de componente utilizado, a ordem de composição mostra efeito relevante tanto em desempenho quanto em estabilidade.

5.5. Por que o *Positional Encoding* Degrada o Modelo

A adição de codificação posicional senoidal ao T→B-*small* reduz o Macro-F1 para $0{,}260$ e produz colapso da classe NOT em parte das sementes. Nesse regime, a combinação entre PE fixo e *embeddings* FastText congelados parece perturbar a geometria original das representações sem oferecer capacidade de adaptação durante o treinamento. Embora o mecanismo exato ainda demande ablação específica, o resultado indica que, neste cenário, atenção sem PE foi consistentemente superior. PE aprendível, escalonado ou aplicado sobre representações contextualizadas permanece como hipótese para trabalho futuro.

5.6. Análise Qualitativa de Erros

Para compreender as diferenças de comportamento entre modelos, foram analisadas 168 instâncias do conjunto de teste nas quais o T→B-*small* classificou corretamente e o BiLSTM-64 errou, considerando a semente 42: 150 textos OFF capturados exclusivamente pelo T→B e 18 textos NOT classificados corretamente apenas pelo T→B.

Entre os OFF capturados exclusivamente pelo T→B, predominam ofensas indiretas, contextuais ou irônicas, sem palavras-chave explicitamente ofensivas. Exemplos representativos incluem “essa *temp* me matou de vergonha alheia”, caracterizada por deboche implícito; “me senti uma idiota (e eu fui)”, marcada por autoironia com carga ofensiva; e “Simplesmente patético tudo isso”, que carrega juízo pejorativo sem xingamento direto. O mecanismo de *self-attention* pondera a coocorrência global de termos aparentemente neutros em posição discriminativa, enquanto a recorrência depende de padrões sequenciais locais e falha quando não há gatilhos lexicais explícitos.

Entre os NOT acertados apenas pelo T→B, observam-se textos com palavras potencialmente ofensivas utilizadas fora de contexto ofensivo. Exemplos incluem “nossa isso deu mó treta na época”, relato neutro sobre conflito passado; “Pobre diabo.”, expressão coloquial sem endereçamento agressivo; e “O insta da racista”, acusação legítima e não ofensa em si. O BiLSTM, dependente de padrões lexicais locais, dispara em palavras isoladas; o T→B, ao ponderar o contexto global, identifica que o tom geral não é ofensivo.

5.7. Limitações

A investigação cobriu um único conjunto de dados, o OLID-BR, o que deixa em aberto a generalização para corpora mais amplos como o ToLD-Br [12], o HateBR [20] ou repositórios agregados de toxicidade em português [14]. Cinco sementes aleatórias por configuração constituem o mínimo prático para *bootstrap* pareado; a ampliação para dez ou mais sementes produziria intervalos de confiança mais estreitos, particularmente úteis

para comparações próximas ao limiar de significância. O *fine-tuning* do BERTimbau utilizou hiperparâmetros padrão da literatura, sem busca em grade sistemática, o que pode subestimar o melhor desempenho possível desse modelo. A análise qualitativa foi conduzida sobre uma única semente, estratégia comum na literatura, mas que deixa margem para variação anedótica. Também não foram avaliados *embeddings* FastText treináveis, *encoders* contextualizados dentro da arquitetura híbrida, nem variantes aprendíveis ou escalonadas de codificação posicional. Por fim, a calibração do limiar de decisão com foco na classe minoritária permanece como hipótese de baixo custo para elevar NOT-F1 sem retreinamento dos modelos.

6. Conclusão e Trabalhos Futuros

Este trabalho apresentou o projeto Curupira, investigação empírica controlada de arquiteturas neurais para detecção de discurso ofensivo em português brasileiro, conduzida sob protocolo metodológico rigoroso, que incluiu separação estrita entre validação e teste, cinco sementes aleatórias por configuração, controle explícito de capacidade paramétrica e teste de significância via *bootstrap* pareado. Cinco achados principais emergem do conjunto experimental e devem ser interpretados no escopo do OLID-BR e do regime de treinamento adotado. Primeiro, o desempenho dos modelos avaliados se concentrou próximo de Macro-F1 0,70, com o BERTimbau superando o melhor modelo leve em 4,3 pontos percentuais. Segundo, a atenção global produziu ganho estatisticamente significativo sobre uma BiLSTM equiparada em capacidade. Terceiro, aumentar a capacidade dos modelos neurais leves não produziu ganho significativo neste cenário. Quarto, a ordem de composição importou, com a variante Transformer→BiLSTM superando a composição inversa em desempenho e estabilidade entre sementes. Quinto, a codificação posicional senoidal somada a *embeddings* congelados degradou substancialmente o desempenho, configurando achado útil para trabalhos futuros com *embeddings* estáticos. O contraste entre o que se observa sob protocolo rigoroso e o que seria obtido em configurações experimentais comuns reforça o chamado de Reimers et al. [17] e Dror et al. [5] pela adoção de múltiplas sementes e testes pareados como padrão em avaliação de modelos de NLP.

Como desdobramentos futuros, quatro direções são consideradas promissoras. A primeira é a agregação dos principais recursos para PT-BR (OLID-BR, ToLD-Br, HateBR e repositórios correlatos) em um corpus unificado, permitindo avaliar se o agrupamento de desempenho observado é específico do OLID-BR ou se persiste em escala maior. A segunda é a substituição do FastText pelo BERTimbau como *encoder* em arquitetura híbrida, combinada com ablação sistemática entre PE aprendível, escalonado e senoidal. A terceira é a calibração do limiar de decisão orientada à classe minoritária, estratégia de baixo custo com potencial de elevar NOT-F1 sem retreinamento. A quarta é a ampliação para dez ou mais sementes por configuração, refinando os intervalos de confiança das comparações fronteiriças. O código e os dados processados serão disponibilizados em repositório público.

Considerações Éticas

Este trabalho utiliza o OLID-BR, conjunto de dados previamente disponibilizado para fins de pesquisa, sem coleta adicional de dados, interação com participantes humanos ou tentativa de reidentificação de usuários. Como a tarefa envolve detecção de discurso ofensivo, reconhece-se que erros de classificação podem gerar impactos relevantes, incluindo

falsos positivos sobre expressões legítimas e falsos negativos sobre conteúdos ofensivos. Assim, os modelos avaliados devem ser interpretados como apoio à análise e não como mecanismo autônomo de decisão em sistemas reais de moderação, os quais exigiriam supervisão humana, auditoria e avaliação contextual de vieses.

Declaração de Uso de IA Generativa

Ferramentas de IA generativa foram utilizadas como apoio auxiliar na pesquisa e na escrita. No projeto Curupira, elas apoiaram decisões de implementação, depuração, melhoria de erros e orientação sobre aspectos de código, sem geração automática do código final; a execução, validação dos resultados e análise permaneceram sob responsabilidade humana. Na escrita, auxiliaram na estruturação do artigo conforme as seções solicitadas, na criação de tópicos-guia, na sugestão de título, resumo e *abstract*, e na adequação de trechos ao formato LaTeX. Todo o conteúdo foi revisado pelos autores para evitar desvios do escopo da pesquisa, inconsistências técnicas ou informações inventadas. As ferramentas utilizadas foram Gemini 3.1 Pro na fase inicial, Claude Code Opus 4.7 com pensamento adaptativo no apoio técnico final, e Claude Opus 4.6 no apoio à escrita.

Referências

[1] Bojanowski, P., Grave, E., Joulin, A. e Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

[2] Davidson, T., Warmsley, D., Macy, M. W. e Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, p. 512–515.

[3] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30.

[4] Devlin, J., Chang, M.-W., Lee, K. e Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, p. 4171–4186.

[5] Dror, R., Baumer, G., Shlomov, S. e Reichart, R. (2018). The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of ACL*, p. 1383–1392.

[6] Efron, B. e Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. CRC Press, Boca Raton.

[7] Fortuna, P. e Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51(4):1–30.

[8] Fortuna, P., Rocha da Silva, J., Soler-Company, J., Wanner, L. e Nunes, S. (2019). A hierarchically-labeled Portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, p. 94–104.

[9] Hochreiter, S. e Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

[10] Huang, Z., Xu, P., Liang, D., Mishra, A. e Xiang, B. (2020). TRANS-BLSTM: Transformer with bidirectional LSTM for language understanding. *arXiv preprint arXiv:2003.07000*.

[11] Khan, S., Fazil, M., Sejwal, V. K., Alshara, M. A., Alotaibi, R. M., Kamal, A. e Baig, A. R. (2022). BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection. *Journal of King Saud University, Computer and Information Sciences*, 34(7):4335–4344.

[12] Leite, J. A., Silva, D. F., Bontcheva, K. e Aker, A. (2020). Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In *Proceedings of AACL-IJCNLP*, p. 914–924.

[13] Malik, J. S., Pang, G. e van den Hengel, A. (2023). Deep learning for hate speech detection: A comparative study. *arXiv preprint arXiv:2202.09517*.

[14] Moreira, L. S., Gibrim, P. T. M., Rocha, L. e Reis, J. C. S. (2026). Anatomy of Data Repositories for the Analysis and Detection of Toxicity in Portuguese. In *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) – Vol. 1*, p. 456–466.

[15] Oliveira, A., Cecote, T., Silva, P., Gertrudes, J., Freitas, V. e Luz, E. (2023). How Good Is ChatGPT For Detecting Hate Speech In Portuguese? In *Proceedings of the 14th Brazilian Symposium in Information and Human Language Technology*, p. 103–112.

[16] Ramos, G., Batista, F., Ribeiro, R., Fialho, P., Moro, S., Fonseca, A., Guerra, R., Carvalho, P., Marques, C. e Silva, C. (2024). A comprehensive review on automatic hate speech detection in the age of the transformer. *Social Network Analysis and Mining*, 14:204.

[17] Reimers, N. e Gurevych, I. (2017). Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of EMNLP*, p. 338–348.

[18] Souza, F., Nogueira, R. e Lotufo, R. (2020). BERTimbau: Pretrained BERT models for Brazilian Portuguese. In *Proceedings of BRACIS*, p. 403–417.

[19] Silva, L. A., Mondal, M., Correa, D., Benevenuto, F. e Weber, I. (2016). Analyzing the targets of hate in online social media. In *Proceedings of the Tenth International AAAI Conference on Web and Social Media*, p. 687–690.

[20] Vargas, F. A., Carvalho, I., Rodrigues de Góes, F., Pardo, T. A. S. e Benevenuto, F. (2022). HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In *Proceedings of LREC*, p. 7174–7183.

[21] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. e Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30*, p. 5998–6008.

[22] Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N. e Kumar, R. (2019). Predicting the type and target of offensive posts in social media. In *Proceedings of NAACL-HLT*, p. 1415–1420.

[23] Trajano, Douglas, Bordini, Rafael H. e Vieira, Renata (2024). OLID-BR: offensive language identification dataset for Brazilian Portuguese. *Language Resources and Evaluation*, 58(4):1263–1289.

CorPop26: Uma atualização das 3.000 Palavras Mais Frequentes do CorPop de 2018

Roges Horacio Grandi[1], **Bianca Pasqualini**[2], **Maria Jose Bocorny Finatto**[2], **Leandro Krug Wives**[1], **Raquel Salcedo Gomes**[1]

[1]PPGIE – CINTED – UFRGS
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

roges.grandi@gmail.com, wives@ufrgs.br, raquel.salcedo@ufrgs.br

[2]PPG-LETRAS-UFRGS – Campus do Vale - Porto Alegre – RS – Brazil

roges.grandi@gmail.com, leandro.wives@ufrgs.br, raquel.salcedo@ufrgs.br

Abstract. *This study aimed to update the list of the 3,000 most frequent words in Brazilian Portuguese based on the CorPop (2018), composed of texts reflecting the average literacy level of readers in the country. The update methodology consisted of comparing the frequencies from original CorPop with those of two other corpora, one based on the Carolina corpus from USP and the other on Wikipedia, thereby expanding the textual base. As a result, 659 words were replaced, representing 21.97% of the original list. Comparative tests across different text analysis tools — NILC Metrix, Vidya Text with original CorPop, and Vidya Text with CorPop26 — demonstrate the influence of the most frequent word list on the results of the Dale and Chall (1995) equation.*

Resumo. *Este estudo teve como objetivo atualizar a lista das 3.000 palavras mais frequentes do português do Brasil com base no CorPop de 2018, formado por textos baseados no letramento médio dos leitores do país. A metodologia de atualização utilizada foi comparar frequências do CorPop original com as de outros dois corpora, um baseado no Carolina da USP e outro na Wikipédia, ampliando a base textual. Como resultado, obteve-se uma renovação de 659 palavras, representando 21,97% das originais. Testes comparativos em diferentes ferramentas de análise textual (NILC Metrix, Vidya Text com CorPop original e Vidya Text com CorPop26) evidenciam a influência da lista de palavras mais frequentes no resultado da fórmula de Dale e Chall de 1995.*

1. Introdução

A fórmula de leiturabilidade de Dale-Chall é um modelo matemático de predição de dificuldade de leitura de um texto que correlaciona uma variável lexical, a *familiaridade vocabular*, a uma variável sintática, o *comprimento médio* das sentenças. Seu modelo original foi desenvolvido em 1948 por Dale e Chall como uma alternativa ao método proposto por Flesch (1948), que se baseia na variável sintática de comprimento médio das sentenças e na variável lexical de número médio de sílabas por palavra. Ambos os modelos foram originalmente desenvolvidos para o inglês norte-americano [Dale and Chall 1948, Flesch 1948].

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 175–187
October 19-22, 2026 ©2026 Association for Computational Linguistics

Dale e Chall (1995) publicaram uma versão revisada de sua fórmula com o objetivo de aprimorar a acurácia preditiva e a adequação a diferentes níveis de leitura. A principal revisão consistiu na atualização da lista de palavras familiares, que passou a refletir de forma mais fiel o vocabulário efetivamente conhecido por estudantes contemporâneos. A lista original, baseada em dados da década de 1940, foi substituída por uma versão revisada e fundamentada em evidências empíricas mais recentes sobre reconhecimento lexical, o que tornou a identificação de palavras frequentes mais precisa [Chall and Dale 1995].

Além disso, a fórmula revisada de Dale e Chall introduziu um fator de ajuste, que adiciona 3,6365 pontos ao índice de legibilidade quando a proporção de palavras "difíceis"(isto é, não frequentes) no texto excede 5%, com a finalidade de compensar a subestimação da dificuldade de ler textos com alta densidade de vocabulário não familiar. A Equação 1 mostra a revisão de Dale e Chall (1995), onde:

- ***PDW (percentage of difficult words)*** = porcentagem de palavras difíceis (não presentes na lista de palavras frequentes)
- ***ASL (average sentence length)*** = quantidade média de palavras por frase.

Conforme apresentado na Equação 1, o cálculo da fórmula revisada de Dale e Chall requer o uso de uma lista de palavras frequentes. Na versão original de 1948, essa lista era composta por aproximadamente 763 palavras familiares. Na versão revisada de 1995, esse conjunto foi expandido para cerca de 3.000 palavras, com o objetivo de melhorar a estimativa da dificuldade lexical [Chall and Dale 1995]. Desde então, a literatura tem adotado esse valor como referência para a equação [DuBay 2004].

$$\textbf{Equação 1: escore bruto} = 0{,}1579 \cdot PDW + 0{,}0496 \cdot ASL + 3{,}6365 \qquad (1)$$

No Brasil, temos um *corpus* do português popular escrito [Pasqualini 2018] composto por textos oriundos de cinco fontes:

1. textos do projeto PorPopular, principalmente do *Diário Gaúcho*[1], mas também de jornais como *Hora de Santa Catarina*[2], voltados às classes C e D, que retratam o leitor médio brasileiro [UFRGS 2016];
2. textos bastante lidos pelo leitor médio brasileiro, de acordo com a pesquisa *Retratos da Leitura no Brasil*[3];
3. textos da coleção *É Só o Começo*, de adaptações de clássicos da literatura brasileira para leitores com baixo letramento[4].
4. textos do jornal *Boca de Rua*[5], produzido por pessoas em situação de rua, com baixa escolaridade e baixo letramento; e

[1]Página inicial: https://diariogaucho.clicrbs.com.br. Acesso em 18 abr. 2026.

[2]Página inicial: https://www.nsctotal.com.br/veiculos/hora. Acesso em 18 abr. 2026.

[3]Página inicial: https://www.prolivro.org.br/pesquisas-retratos-da-leitura/as-pesquisas-2. Acesso em 18 abr. 2026.

[4]Essa coleção foi criada a partir do projeto de valorização do livro, uma iniciativa do Serviço Social da Indústria (SESI), com incentivos e auxílio da Unesco e do Programa Brasil Alfabetizado, do Ministério da Educação e Cultura (MEC) [Vieira 2010].

[5]Página inicial: https://jornalbocaderua.wordpress.com. Acesso em: 18 abr. 2026.

5. textos do *Diário da Causa Operária*[6], imprensa operária brasileira produzida também por pessoas dentro da faixa média de letramento do país.

A partir desse *corpus*, compilado de acordo com o nível médio de letramento dos brasileiros, [Pasqualini 2018] criou uma lista ordenada de frequência de 9.904 palavras, da qual se pode extrair a lista das 3.000 palavras mais frequentes e usá-la nas equações de Dale e Chall no português do Brasil.

Este estudo propõe uma reordenação das 3.000 palavras mais frequentes do CorPop por meio de uma investigação cruzada com palavras frequentes encontradas em dois *corpora*:

1. ***Corpus* Carol Pop BR**: derivado de uma seleção de textos do *corpus* Carolina[7] da Universidade de São Paulo (USP), do qual se procurou excluir textos de vocabulário específico e textos com origem em Portugal, mais textos do Diário Gaúcho [Grandi 2026a].
2. ***Corpus* Wikipédia PTLimpo**: palavras extraídas de artigos da Wikipédia em português [Grandi 2026b].

Tanto o *Corpus* Carol Pop BR quanto o *Corpus* Wikipédia PTLimpo foram processados para extrair suas palavras mais frequentes, por meio de um processamento automatizado de remoção de tremas, exclusão de sequências de caracteres formadas somente por números romanos e separação de nomes. As palavras restantes foram as selecionadas para computar frequências, das quais foram selecionadas as 10.000 mais frequentes. Por fim, foi realizada uma análise visual dos três *corpora*: CorPop, Carol Pop BR e Wikipédia PTLimpo, para eliminar abreviaturas, letras soltas remanescentes, como "b"e "s", palavras em português não pertencentes à grafia brasileira, como "património", e estrangeirismos não incorporados ao português do Brasil, conforme o Vocabulário Ortográfico Comum da Língua Portuguesa (VOC[8]). Chamamos as palavras remanescentes ao final desse processo de *CarolPopBR10mil* e *WikipédiaPTLimpo10mil*.

A atualização de frequências do CorPop original para formar o CorPop26 foi realizada por meio da leitura, palavra por palavra, do CorPop, em ordem de frequência, das palavras mais frequentes às menos frequentes. Se a palavra lida estivesse simultaneamente no *CarolPopBR10mil* e no *WikipédiaPTLimpo10mil*, a palavra era selecionada para formar a lista de palavras mais frequentes do CorPop26. Essa etapa foi necessária para garantir a evolução do CorPop original, evitando vieses muito intensos do vocabulário jornalístico da imprensa paulista presente no *corpus* Carolina e do vocabulário lusitano também bastante presente no *corpus* original da Wikipédia.

Assim, este artigo foi organizado da seguinte forma: a seção *Fundamentação Teórica* discorre sobre leiturabilidade, frequência de palavras e representatividade de textos em *corpora*. A seção *Metodologia* relata o processo de renovação adotado. A seção *Processo de Criação do CorPop26* detalha a criação dos *corpora* Carol Pop BR e Wikipédia PTLimpo, assim como critérios e os passos de seleção das 3.000 palavras do CorPop26. A seção *Resultados* comenta um teste em que se comparam resultados de Dale-Chall obtidos em diferentes plataformas de análise de leiturabilidade de textos em português do

[6]Página inicial: https://causaoperaria.org.br. Acesso em 18 abr. 2026.
[7]Página inicial: https://sites.usp.br/corpuscarolina. Acesso em 18 abr. 2026.
[8]Página inicial: https://voc.cplp.org/index.php?action=von&csl=br. Acesso em 18 abr. 2026.

Brasil. Concluiu-se que as métricas de Dale-Chall que adotam a lista de 3.000 palavras selecionadas do CorPop26 apresentam bons níveis de acurácia em relação aos níveis de escolaridade estimados por esse método.

2. Fundamentação Teórica

Finatto e Paraguassu (2021) definem leiturabilidade como a potencial facilidade ou dificuldade de leitura de um texto, determinada por fatores linguísticos (referentes a escolhas lexicais e sintáticas) que, por sua vez, estão relacionados ao perfil de leitor pretendido do texto, que pode ser alguém com maior ou menor escolaridade, com mais ou menos conhecimento prévio do assunto do texto [Finatto and Paraguassu 2021], entre outros aspectos. Estimativas de leiturabilidade têm sido tradicionalmente modeladas por meio do uso de variáveis sintáticas e lexicais, conforme proposto por pesquisadores como Flesch[9], Dale, Chall e outros [Flesch 1948, Dale and Chall 1948, Chall and Dale 1995, DuBay 2004]. O domínio das palavras mais frequentes de um idioma é fundamental para a compreensão textual, o que reforça a relevância das listas de frequência e dos modelos de leiturabilidade, como os propostos por Dale e Chall [Dale and Chall 1948, Chall and Dale 1995, Nation 2001].

Em relação à validade de análises textuais, a Linguística de *Corpus* enfatiza que a validade de análises linguísticas depende diretamente da representatividade dos dados utilizados. Biber (1993) argumenta que os *corpora* devem refletir a diversidade de usos da língua, enquanto McEnery e Hardie (2012) destacam que vieses de domínio podem comprometer generalizações linguísticas [McEnery and Hardie 2011]. Nesse sentido, a combinação de vários *corpora* constitui uma estratégia relevante para mitigar vieses e ampliar a cobertura lexical.

3. Metodologia

Idealmente, a construção de listas de frequência lexical deve buscar minimizar vieses amostrais. Entre os critérios fundamentais, destaca-se a abrangência regional do *corpus* utilizado. Assim, no caso do português do Brasil, recomenda-se a adoção de *corpora* de cobertura nacional, constituídos por textos produzidos em português brasileiro, de modo a assegurar maior representatividade linguística.

Para a estimação da leiturabilidade por faixa etária, é igualmente desejável a disponibilidade de textos anotados por nível de ensino, a fim de viabilizar a condução de testes de acurácia. Idealmente, tais dados devem apresentar abrangência nacional, contemplando diferentes regiões do Brasil, de modo a assegurar maior representatividade educacional e linguística. No caso específico da fórmula de Dale e Chall (1995), se a classificação dos textos por nível de ensino não estiver de acordo com a realidade brasileira, é possível, a partir de um conjunto de textos representativos, ajustar os parâmetros da equação, para minimizar a margem de erro. Um desses parâmetros refere-se ao tamanho da lista de palavras frequentes, cujo valor de referência na literatura é de aproximadamente 3.000 vocábulos.

Diante do exposto, na ausência de *corpora* ideais, optou-se por atualizar a lista das 3.000 palavras mais frequentes do CorPop, tomando como base sua compilação de 9.940

[9] [Martins et al. 1996] adaptaram a fórmula de Flesch para o português do Brasil.

ocorrências, devido à representatividade sociolinguística e do nível médio de letramento da população e registro do português popular escrito. Para a atualização, selecionaram-se *corpora* de textos em língua portuguesa, preferencialmente do português do Brasil, buscando-se a maior abrangência possível e evitando-se vocabulários especializados. Esses *corpora* foram utilizados para o cruzamento de frequências, com o objetivo de reordenar as frequências observadas no CorPop original. A estratégia de reordenação consistiu na iteração sobre o léxico do CorPop original, verificando-se, para cada palavra, sua ocorrência nos demais *corpora* selecionados. Quando identificada uma intersecção, o item foi mantido na lista final até que se alcançasse o conjunto de 3.000 palavras.

4. Processo de Criação do CorPop26

Com base na metodologia apresentada, buscaram-se *corpora* abrangentes e bem elaborados para realizar a intersecção com as palavras contabilizadas no CorPop original. Dentre os *corpora* disponíveis para uso acadêmico, foram encontrados o *Brazilian Portuguese Web as Corpus* (BrWaC[10]), o *Corpus* Geral do Português Brasileiro Contemporâneo (Carolina), um *corpus* do *Diário Gaúcho*, um *corpus* da Wikipédia e um *corpus* da Google, o mC4[11].

4.1. *Corpus* BrWac

O BrWac é um *corpus* que foi disponibilizado publicamente pelo *Neurocognition and Natural Language Processing Research Lab* do Instituto de Informática da Universidade Federal do Rio Grande do Sul (UFRGS) em 2017, composto por 3,53 milhões de documentos e 2,68 bilhões de *tokens*. Embora seja bastante extenso, não foi possível extrair adequadamente as palavras desse *corpus*, pois o arquivo apresenta muitos problemas de codificação de caracteres decorrentes da concatenação dos milhões de documentos que o compõem, o que inviabilizou seu uso.

4.2. *Corpus* Carolina

O Carolina é um *corpus* desenvolvido pela Universidade de São Paulo (USP) contendo textos em português coletados de 1970 a 2021 [USP 2023]. Neste estudo, foi utilizada a versão de 2023, que contém mais de dois milhões de textos e 823 milhões de *tokens*. Com a finalidade de minimizar vieses, os *corpora* e *datasets* do Carolina listados na Tabela 1 não foram utilizados.

Tabela 1. Análise de Corpora e Datasets Desconsiderados

Corpus / Dataset	Características	Motivo
Brazilian Stock Market Tweets with Emotions	*Corpus* para análise de emoções em *tweets* sobre a Bolsa de Valores de São Paulo em 2018 [UNICAMP 2018].	Terminologia especializada do mercado financeiro.

[10]Página inicial: https://huggingface.co/datasets/UFRGS/brwac. Acesso em 19 abr 2026.
[11]Página inicial: https://aaas.blog/dataset/mc4-dataset. Acesso em 27 abr. 2026.

Corpus / Dataset	Características	Motivo
Datasets of Neuropsychological Language Tests in Brazilian Portuguese	*Datasets* de neuropsicologia elaborados da Pontifícia Universidade Católica do Rio Grande do Sul.	Terminologia especializada da área médica [Casanova et al. 2025].
HAREM	*Corpora* mantidos pela Linguateca[12].	Forte influência do português europeu [Santos et al. 2015].
Leg2Kids	*Corpus* de 36.413 legendas de filmes e séries dos gêneros "família"e "animação".	Vocabulário direcionado a crianças [Hartmann 2019].
Portuguese Tweets for Sentiment Analysis	*Dataset* de 235,3 MB de *tweets* voltados às análises de polaridade de sentimentos.	Muitos dos *tweets* têm erros de grafia, abreviaturas e jargões de internet.
Revista Pesquisa FAPESP *Parallel Corpora*	Link informado no Carolina[13] está quebrado.	Impossibilidade de acesso.
Judiciário	*Corpora* do STF com 38.068 textos.	Vocabulário especializado do Direito.
Legislativo	*Corpora* do Senado brasileiro constituídos por 14 textos.	Vocabulário especializado do Poder Legislativo.
Social Media	O link informado no Carolina[14] leva a uma página sem opções para *download*.	Impossibilidade de acesso.
Wikis	*Corpora* formados por 960.139 textos da Fundação Wikimedia.	O *corpus* completo de artigos da Wikipédia em português utilizado neste estudo tornou esta coleção desnecessária.

4.3. *Corpus* Bruto do *Diário Gaúcho*

Também foi utilizado o *corpus* bruto do *Diário Gaúcho*, de 2008, disponibilizado pelo projeto PorPopular da UFRGS [UFRGS 2016]. Esse *corpus* reúne textos coletados ao longo de 116 dias daquele ano, totalizando 2,89 MB. Nesse mesmo *site*, há um arquivo de 2010 com extrações de textos do mesmo periódico. Todavia, esse arquivo não pode ser aproveitado devido a problemas de codificação de caracteres.

4.4. Formação do Corpus Carol Pop BR

O *Corpus* Carol Pop BR [Grandi 2026a] foi formado pela unificação dos textos selecionados do Carolina aos textos selecionados do *Diário Gaúcho*, conforme descrito anteriormente. Após a concatenação dos textos em um *corpus* unificado com 510 MB e 90,1 milhões de *tokens*, foi executada uma rotina de remoção de tremas.

[12]Centro de pesquisa e desenvolvimento, repositório de *corpora*, de ferramentas e recursos linguísticos e um polo de articulação científica entre pesquisadores de países lusófonos.

[13]Página: http://www.nilc.icmc.usp.br/nilc/tools/Fapesp%20Corpora.htm. Acesso em 19 abr. 2026.

[14]Página inicial: https://www.wattpad.com. Acesso em 19 abr. 2026.

4.5. Formação do Corpus Wikipédia PTLimpo

Além da compilação de textos que formou o Corpus Carol Pop BR, buscou-se uma coleção de textos com foco em abrangência. As opções encontradas foram:

1. a coleção de todos os artigos da Wikipédia em português, disponibilizada em seu repositório oficial[15], com mais de um milhão de artigos, os quais ocupam 7,2 GB após descompactados; e
2. o extrato em língua portuguesa do *Multilingual Colossal Clean Crawled Corpus* (mC4), um *corpus* textual multilíngue de 27 TB desenvolvido pela Google para treinar modelos de linguagem. O extrato em português do mC4 tem aproximadamente 524 GB.

Analisadas as alternativas, optou-se pela utilização de artigos da Wikipédia, uma vez que esses textos passam por revisão colaborativa contínua, com tendência à clareza e à padronização linguística, ainda que apresentem variação quanto ao nível de complexidade. A extração de palavras do Corpus Wikipédia PTLimpo foi realizada em Python, com o módulo Gensim da biblioteca Wikicorpus[16]. Para essa extração, letras maiúsculas e minúsculas foram preservadas, bem como palavras compostas por uma única letra. Finalizando o tratamento, tanto no Corpus Carol Pop BR como no Corpus Wikipédia PTLimpo foi executada uma rotina de remoção de tremas.

4.6. Extração das 10 mil Palavras mais Frequentes dos *Corpora* Carol Pop BR e Wikipédia PTLimpo

Os *corpora* Carol Pop BR e Wikipédia PTLimpo passaram pela mesma sequência de linhas de comando em um sistema operacional Linux para extrair suas 10 mil palavras mais frequentes.

4.6.1. Extração de Palavras

Para extrair as palavras dos *corpora*, foi executado o seguinte comando, onde `$CORPUS_FILE` é o *corpus* Carol Pop BR ou o Wikipédia PTLimpo. Foi usada uma expressão regular no comando `grep` para extrair todas as sequências contínuas de caracteres alfabéticos do *corpus*, retornando cada ocorrência como um *token* individual em uma nova linha. Em seguida, esses *tokens* foram direcionados por meio de um pipe para o comando `sort`, que ordenou os termos alfabeticamente. Por fim, o resultado foi armazenado no arquivo `$ORDENADAS`:

```
grep -oE
    "[[:alpha:]]
    "$CORPUS_FILE" | sort --parallel=4 > "$ORDENADAS"
```

Para liberar espaço, o `$CORPUS_FILE` foi removido. A etapa seguinte foi aplicar o comando `uniq` com a opção -c sobre os arquivos previamente ordenados de *tokens*, com o objetivo de agrupar ocorrências consecutivas idênticas e contabilizar suas frequências. Desse modo, para cada palavra, foi gerada uma linha contendo o número de ocorrências seguido do respectivo *token*:

[15] Página inicial: https://dumps.wikimedia.org/ptwiki/latest. Acesso em 19 abr. 2026.

[16] Página inicial: https://github.com/piskvorky/gensim/tree/develop/gensim. Acesso em 19 abr. 2026.

```
uniq -c "$ORDENADAS" > "$CONTADAS"
```

A próxima etapa foi criar dois arquivos, o `$COMUNS_FREQ` e o `$NOMES_FREQ`, por meio do comando `awk`, tendo como parâmetro de entrada o arquivo `$CONTADAS` de cada *corpus*. Inicialmente, cada palavra foi normalizada para minúsculas para agrupar variações de capitalização sob a mesma chave, assim como suas frequências de ocorrência. Em seguida, foram descartadas palavras com frequência inferior a 20 e calculou-se a proporção de usos com letra maiúscula. Palavras cuja escrita com inicial maiúscula era superior a 95% foram gravadas no arquivo `$NOMES_FREQ`. As demais foram consideradas comuns e armazenadas no arquivo `$COMUNS_FREQ`. Por fim, foi utilizado o comando *head* para se obter as 10 mil palavras mais frequentes dos respectivos arquivos `$COMUNS_FREQ`:

```
head -n 10000 "$COMUNS_FREQ" > "$TOP10K"
```

O último trabalho realizado nessa etapa foi remover visualmente, não apenas nos *corpora* Carol Pop BR e Wikipédia PTLimpo, mas também na lista de frequências do CorPop original, abreviaturas, letras soltas, como "b"e "s", estrangeirismos e palavras cuja grafia não corresponde à do português do Brasil, conforme o VOC.

4.7. Finalização da Criação do CorPop26

A última etapa da criação do CorPop26 consistiu em uma análise comparativa entre os vocábulos do CorPop original e os vocábulos dos *corpora* Carol Pop BR e Wikipédia PTLimpo. O CorPop original continuou sendo a base, como afirmamos anteriormente, por sua representatividade sociolinguística e registro do português popular escrito. Observa-se que o *corpus* Carol Pop BR, por ser composto predominantemente por textos do *Jornal Folha de São Paulo* (97,2%), tem um forte viés do vocabulário jornalístico paulista. Os artigos em língua portuguesa da Wikipédia, por sua vez, componentes do *corpus* Wikipédia PTLimpo, apesar de serem escritos, em sua maioria, por brasileiros (em torno de 78%), ainda têm um viés considerável do português lusitano.

Isto posto, uma intersecção de palavras dos três *corpora* ajudou a diminuir os vieses dos *corpora* Carol Pop BR e Wikipédia PTLimpo. O resultado foi o CorPop26, um conjunto das 3.000 palavras mais frequentes do português do Brasil, com base no CorPop original, mas influenciado pelos *corpora* Carol Pop BR e Wikipédia PTLimpo para uma ordenação que reflita uma abrangência maior, tanto textual como regional. A atualização das palavras mais frequentes resultou na substituição de 659 vocábulos, correspondendo a 21,97% do conjunto original de 3.000 itens[17].

5. Resultados

Para avaliar o impacto da atualização do vocabulário, foram realizados testes comparativos com três configurações distintas da fórmula de Dale-Chall (1995). Com base na pontuação bruta resultante da aplicação da equação, é possível estimar tanto o número de anos de escolarização (contados a partir do primeiro ano do Ensino Fundamental) quanto o respectivo nível de escolaridade, conforme apresentado na Tabela 2 [Paraguassu 2024, p. 174].

[17] A lista das palavras substituídas está disponível em [Grandi and Pasqualini 2026].

Tabela 2. Interpretação de Dale-Chall (1995) por nível de escolaridade

Pontuação	Anos	Escolaridade
< 5	≤ 4	Nível 1: até o 4º ano do Ensino Fundamental
5 a < 6	5 a 6	Nível 2: do 5º ao 6º ano do Ensino Fundamental
6 a < 7	7 a 8	Nível 3: do 7º ao 8º ano do Ensino Fundamental
7 a < 8	9 a 10	Nível 4: do 9º ano do Ensino Fundamental ao 1º ano do Ensino Médio
8 a < 9	11 a 12	Nível 5: do 2º ano do Ensino Médio aos primeiros anos do Superior
9 a < 10	13 a 14	Nível 6: Ensino Superior
≥ 10	≥ 15	Nível 7: Ensino Superior e Pós-graduação

A Tabela 3 mostra um comparativo entre os resultados obtidos utilizando-se o sistema NILC-Metrix [Leal et al. 2023] e a Vidya Text [Grandi et al. 2026] em duas configurações: uma com o vocabulário do CorPop e outra com o vocabulário do CorPop26. Foram selecionados cinco trechos de caráter institucional, produzidos por órgãos de saúde e cidadania e voltados à comunicação direta com o cidadão. Conforme detalhado nessa Tabela, embora as ferramentas usem a mesma equação [Chall and Dale 1995], os resultados indicam que as variações nas listas de palavras mais frequentes influenciam a métrica final. Essa sensibilidade demonstra que a escolha da base de referência é uma variável importante e pode modificar significativamente os índices de legibilidade e a classificação do nível de complexidade dos textos. A interpretação da pontuação dos textos analisados na Tabela 3 pode ser consultada na Tabela 2.

A atualização lexical é refletida especialmente nos resultados das amostras 2 e 3, em que houve reclassificação do nível dos textos. Essa mudança ocorreu porque termos como "oportunidade", "atendimento"e "respeito"(em azul) passaram a constar do CorPop26. Vê-se, assim, que a nova base de referência oferece uma estimativa de leiturabilidade voltada à realidade do português brasileiro contemporâneo. Na amostra 5, observa-se uma divergência significativa de 2,3 pontos entre os resultados: o NILC-Metrix classificou o texto como Nível 5 (8,5 pontos), ao passo que a Vidya Text o classificou como Nível 3 (6,2 pontos). Essa diferença exemplifica o impacto direto da escolha da base textual na sensibilidade da ferramenta. Enquanto o NILC-Metrix tende a elevar os índices de dificuldade de itens lexicais que, embora de fácil compreensão intuitiva (como "decisões"e "apoiada", em vermelho), não estão em sua base de referência, o CorPop26 reflete uma seleção lexical mais próxima da realidade do letramento brasileiro atual, como mencionado anteriormente.

6. Considerações Finais e Trabalhos Futuros

Diante do objetivo de atualizar as listas das 3.000 palavras mais frequentes do CorPop original de 2018, tendo como metodologia a ampliação da base textual e mantendo o CorPop como *corpus* de referência (e considerando-se que ele é um bom indicativo do nível de letramento médio dos brasileiros), obteve-se uma renovação de 659 (21,97%) das palavras selecionadas. Os testes com a fórmula de Dale-Chall (1995), em três configurações (NILC-Metrix, Vidya Text com o CorPop e Vidya Text com CorPop26), indicam que a lista de frequências influencia o resultado. A ampliação da base textual usada para renovar a lista de palavras mais frequentes, utilizada neste estudo, contribui para sua representa-

Tabela 3. Comparativo de Índices de Legibilidade por Ferramenta

Textos de amostra	FC1	FC2	NILC
(1) O cultivo de bons hábitos alimentares, a prática regular de exercícios físicos, a boa qualidade do sono e o acompanhamento médico periódico são fatores que atuam em conjunto para preservar o bem-estar físico e mental.	Nv: 6 (10,8 pts)	Nv: 6 (9,4 pts)	Nv: 6 (9,89 pts)
(2) A vacinação é essencial para manter a criança saudável. Durante a triagem neonatal, deve-se aproveitar a oportunidade para avaliar se a criança recebeu as vacinas indicadas ao nascer ou se há necessidade de atualizá-las.	Nv: 6 (9,1 pts)	Nv: 5 (8,2 pts)	Nv: 6 (10,05 pts)
(3) O SUS oferece cuidado em saúde para todas as faixas etárias, incluindo exames preventivos, planejamento familiar, atendimento pré-natal e parto. Você tem direito a ser tratado com respeito em todos os serviços.	Nv: 5 (8,4 pts)	Nv: 4 (7,6 pts)	Nv: 6 (9,36 pts)
(4) O Programa Nacional de Saneamento Rural estabelece diretrizes para ações em áreas rurais, visando à universalização do acesso. Entretanto, pouco foi implementado com base em sua orientação para garantir os direitos humanos.	Nv: 5 (8,4 pts)	Nv: 5 (8,4 pts)	Nv: 6 (10,35 pts)
(5) Se a pessoa consegue tomar decisões, dizer o que quer e o que não quer, ela não precisa de um curador (alguém que decide no seu lugar); ela precisa apenas de ajuda. Para isso, a Tomada de Decisão Apoiada é suficiente.	Nv: 3 (6,2 pts)	Nv: 3 (6,2 pts)	Nv: 5 (8,50 pts)

Legenda: FC1 = Vidya Text com CorPop; FC2 = Vidya Text com CorPop26; Nv = nível; pts = pontos. Nível 6: dos últimos anos do Ensino Superior à Pós-graduação; Nível 5: do 2º ano do Ensino Médio aos primeiros anos do Ensino Superior; Nível 3: do 7º ao 8º ano do Ensino Fundamental. Palavras em azul constam no CorPop26 (FC2); palavras em vermelho permanecem classificadas como complexas.

Fontes das amostras: (1) [Hospital Português 2021]; (2) [Brasil. Ministério da Saúde 2023]; (3) [Brasil. Ministério da Saúde 2024]; (4) [Aguiar et al. 2025]; (5) [Ceará. Defensoria Pública do Estado 2024]

tividade. Ressalta-se, no entanto, a dificuldade de se obter uma "lista definitiva"frente às centenas de milhares de palavras que compõem a língua portuguesa no Brasil, conforme registra o Vocabulário Ortográfico Comum da Língua Portuguesa (VOC).

Os trabalhos futuros, em curto e médio prazo, incluem a realização de outros testes com o método de Dale-Chall (1995), em diferentes contextos. Sendo a língua dinâmica, e considerando-se a diversidade de *corpora* existentes, certamente outros *corpora* poderão gerar listas de frequências que reflitam com mais fidelidade o português escrito no cotidiano brasileiro, desde que tenham quantidade e diversidade de tipos textuais empregados no Brasil e que sejam tomados os devidos cuidados para minimizar vieses.

7. Agradecimentos

Esta pesquisa foi apoiada pelo CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), processo PQ 307088/2023-5, pelos Institutos Nacionais de Ciência e Tecnologia – CNPq-TILD-IAR (processo 408490/2024-1), pela PROBIC FAPERGS-UFRGS e pela FAPERGS Ed. 06/2025 – PROGRAMA PESQUISADOR GAÚCHO –

PqG (processo 25/2551-0002702).

8. Declaração de Ética

Este estudo utilizou dados linguísticos de domínio público e corpora acadêmicos de prestígio (CorPop, *Corpus* Carolina e Wikipédia). Não houve envolvimento de seres humanos, de prontuários médicos ou de dados pessoais sensíveis no desenvolvimento das listas de palavras nem nos testes comparativos. A pesquisa segue os padrões éticos da Linguística Computacional, buscando melhorar ferramentas de acessibilidade na comunicação em saúde.

9. Limitações

A principal limitação deste estudo é o dinamismo da língua portuguesa; qualquer lista fixa de palavras frequentes necessita de atualizações regulares para manter sua relevância. Apesar do uso de três *corpora* diferentes ter ajudado a reduzir vieses regionais e estilísticos, o CorPop26 criado é uma base de linguagem geral e pode precisar de ajustes adicionais ao ser usado em áreas técnicas muito específicas.

10. Declaração sobre o Uso de IA Generativa

Em conformidade com a Portaria CNPq 2.664/2026, a versão 0.38.2 do Gemini CLI foi usada para o refinamento estilístico, organização estrutural e processamento de contagens lexicais específicas do artigo durante a preparação deste manuscrito. Nenhuma ferramenta de IA foi utilizada para gerar os dados de frequência, nem para a metodologia central de atualização da lista de palavras.

Referências

Aguiar, A. M. S. et al. (2025). *Experiências sobre água e saneamento rural na América Latina e no Caribe*. Instituto René Rachou, Fiocruz Minas, Belo Horizonte. Amostra de texto 4.

Brasil. Ministério da Saúde (2023). *Caderneta da criança: passaporte da cidadania (menina)*. Ministério da Saúde, Brasília, 6 edition. Amostra de texto 2.

Brasil. Ministério da Saúde (2024). *Conheça seus direitos no SUS! (Guia para migrantes e refugiados)*. Ministério da Saúde, Brasília. Amostra de texto 3.

Casanova, E., Treviso, M., Hübner, L. C., and Mansur, L. L. (2025). Dnlt-bp: Dataset para avaliação de complexidade textual em português brasileiro. `https://github.com/nilc-nlp/DNLT-BP`. Dataset. Acesso em: 5 jan. 2026.

Ceará. Defensoria Pública do Estado (2024). *Cidadania sem Barreiras: curatela e tomada de decisão apoiada*. Defensoria Pública do Estado do Ceará, Fortaleza. Amostra de texto 5.

Chall, J. S. and Dale, E. (1995). *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books, Cambridge.

Dale, E. and Chall, J. S. (1948). A formula for predicting readability: Instructions. *Educational Research Bulletin*, pages 37–54.

DuBay, W. H. (2004). *The Principles of Readability*. Impact Information, Costa Mesa.

Finatto, M. J. B. and Paraguassu, L. B. (2021). *Acessibilidade textual e terminológica*. EDUFU.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Grandi, R. H. (2026a). Corpus Carol Pop BR. Kaggle. Dataset.

Grandi, R. H. (2026b). Corpus wikipédia PT limpo. Kaggle. Dataset.

Grandi, R. H. et al. (2026). Vidya text: an educational tool to support the teaching of writing and the study of brazilian portuguese vocabulary. https://vidyatext.nuvem.ufrgs.br. Acesso em: 19 jul. 2026.

Grandi, R. H. and Pasqualini, B. (2026). CorPop 2018/2026. Kaggle. Dataset.

Hartmann, N. S. (2019). Leg2kids. `https://sites.google.com/view/nilc-usp/leg2kids`. Corpus. Acesso em: 5 jan. 2026.

Hospital Português (2021). *Check-up da pessoa idosa: acompanhamento médico e exames de rotina*. Salvador. Amostra de texto 1. Informativo institucional.

Leal, S. E., Duran, M. S., Scarton, C. E., Hartmann, N. S., and Aluísio, S. M. (2023). Nilc-metrix: assessing the complexity of written and spoken language in brazilian portuguese. *Language Resources and Evaluation*.

Martins, T. B. F. et al. (1996). Readability formulas applied to textbooks in brazilian portuguese. Technical report, ICMC-USP.

McEnery, T. and Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

Nation, I. S. P. (2001). *Learning Vocabulary in Another Language*. Cambridge University Press, Cambridge.

Paraguassu, L. B. (2024). *Linguagem Simples na Saúde: contribuições dos estudos da linguagem para o Letramento em Saúde no Brasil*. Tese (doutorado em letras), Universidade Federal do Rio Grande do Sul, Porto Alegre.

Pasqualini, B. (2018). Corpop: Um corpus do português popular escrito. Tese, Programa de Pós-Graduação em Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre.

Santos, D., Cardoso, N., Sarmento, Luís e Rocha, P., and Vieira, R. (2015). Harem. `https://www.linguateca.pt/HAREM`. Corpus. Lisboa: Linguateca. Acesso em: 5 jan. 2026.

UFRGS (2016). Porpopular: Léxico do português popular brasileiro. `http://www.ufrgs.br/textecc/porlexbras/porpopular/index.php`. Acesso em: 18 abr. 2026.

UNICAMP (2018). Brazilian stock market tweets with emotions. `https://www.kaggle.com/datasets/fernandojvdasilva/stock-tweets-ptbr-emotions`. Dataset. Acesso em: 5 jan. 2026.

USP (2023). Carolina: Corpus geral do português brasileiro contemporâneo. `https://sites.usp.br/corpuscarolina/repositorios-2023`. Corpus. Acesso em: 5 jan. 2025.

Vieira, G. d. O. (2010). Adaptação para novos leitores: como a literatura clássica adaptada fornecida às escolas do ensino público e utilizada pelos professores no processo de ensino estimula a leitura de obras originais.

Vidya Syl: A syllable division algorithm for Brazilian Portuguese adhering to the Common Orthographic Vocabulary of the Portuguese Language

Roges Horacio Grandi[1], Leandro Krug Wives[1], Maria Jose Bocorny Finatto[2], Raquel Salcedo Gomes[1], Bianca Pasqualini[2]

[1]PPGIE – CINTED – UFRGS
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

roges.grandi@gmail.com, leandro.wives@ufrgs.br, raquel.salcedo@ufrgs.br

[2]PPG-LETRAS-UFRGS – Campus do Vale - Porto Alegre – RS – Brazil

maria.finatto@ufrgs.br, bianca.pasqualini@gmail.com

Abstract. *Syllabification is a fundamental task in natural language processing (NLP) applications, particularly in readability assessment through metrics such as the Flesch Index. This paper presents Vidya Syl, a new syllabification algorithm implemented in Java, developed in compliance with the Brazilian section of the Common Orthographic Vocabulary of the Portuguese Language, known as VOC. The methodological approach consisted of augmenting a rule-based algorithm grounded in phonological and graphical rules with morphological and orthographic rules, until it became compliant with VOC. The experimental results demonstrate improvements in both accuracy and lexical coverage.*

Resumo. *A divisão silábica é uma tarefa fundamental em aplicações de processamento de linguagem natural (PLN), especialmente na avaliação de leiturabilidade por meio de métricas como o Índice Flesch. Este artigo apresenta o Vidya Syl, um novo algoritmo de divisão silábica implementado em Java, desenvolvido em conformidade com a seção brasileira do Vocabulário Ortográfico Comum da Língua Portuguesa, denominada VOC. A abordagem metodológica adotada consistiu em incrementar, com regras morfológicas e ortográficas, um algoritmo baseado em regras fonológicas e gráficas, até que se tornasse aderente ao VOC. Os resultados experimentais demonstram melhorias tanto em acurácia quanto em cobertura lexical.*

1. Introduction

Syllable division consists of segmenting words into phonological units called syllables and is essential for applications such as hyphenation, speech synthesis, and reading instruction. In Portuguese, this task is guided by phonological and orthographic principles, such as the mandatory presence of a vowel in each syllable and the specific treatment of vowel and consonant clusters [Câmara Jr. 1992]. Syllable division is a central task in natural language processing (NLP) applications, such as tasks involving estimating the readability of texts using the Flesch index, for example [Flesch 1948].

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 188–199
October 19-22, 2026 ©2026 Association for Computational Linguistics

Despite the existence of rule-based algorithms, many exhibit limitations in covering exceptions and in normative adherence. In this context, the Common Orthographic Vocabulary of the Portuguese Language (VOC) emerges as a standardized lexical resource that enables the validation of syllable divisions according to the official orthography [International Institute of the Portuguese Language 2024].

VOC is the platform that houses the instruments that legally determine the Portuguese orthography. It was officially recognized by the Member States of the Community of Portuguese Language Countries (CPLP) in the final conclusions of the X Conference of Heads of State and Government of the CPLP, held in July 2014 in Dili. In the VOC, it is possible to access the specific dictionary of each member country and to check the frequencies of different forms and syllable divisions used in each country. This work proposes an algorithm that initially relied on phonological rules and progressively incorporated morphological and orthographic constraints, ultimately achieving compliance with the Brazilian section of VOC, namely the *Vocabulário Ortográfico da Língua Portuguesa (VOLP)*, coordinated by Evanildo Bechara [International Institute of the Portuguese Language 2026].

Sections 2 through 7 are organized as follows. Section 2 reviews related work, highlighting rule-based approaches to syllable division in Brazilian Portuguese and discussing their contributions and limitations. Chapter 3 introduces the VOC and its Brazilian section VOLP, which serves as the normative lexical basis for validating syllabification. Chapter 4 presents the Vidya Syl algorithm, detailing its hybrid design, architecture, and innovations aimed at improving adherence to phonological and orthographic rules. Section 5 describes the methodology, including large-scale data extraction from the VOLP, preprocessing steps, baseline evaluation, and incremental rule refinement. Section 6 reports the results, demonstrating the high level of alignment achieved by the proposed method. Finally, Section 7 concludes the study and outlines future research directions.

2. Related Works

Silva, Braga, and Resende Jr. (2008) proposed an approach combining two algorithms based on linguistic rules for Brazilian Portuguese: one for syllabification (with 20 rules) and another for stress assignment (with 19 rules), both aimed at applications such as text-to-speech systems. The methods analyze word structure based on vowels and grapheme-to-phoneme context, prioritizing phonological and orthographic principles rather than dictionary-based approaches, which reduces memory usage and allows handling of new words [Silva et al. 2008].

Rocha and Neto (2013) present an algorithm for syllabification in Portuguese that is directly based on the work of Silva, Braga, and Resende Jr. (2008), extending and refining the rule-based approach. The method maintains a focus on phonological and orthographic principles but introduces adjustments to better handle problematic cases, such as consonant clusters, digraphs, and language-specific exceptions. Experimentally evaluated, the algorithm demonstrates good performance and robustness, reinforcing the viability of rule-based approaches for linguistic processing, especially in scenarios that require low computational cost and independence from dictionaries [Rocha and Neto 2013].

Amorim (2024) proposed a rule-based syllabification algorithm for Portuguese implemented in the syllable-pt library, explicitly building upon the works of Silva, Braga, and Resende Jr. (2008) and Rocha and Neto (2013). The approach preserves the linguistic-rule foundation of previous studies while incorporating additional refinements to improve syllable splitting in practical applications [Amorim 2024].

Implemented as a lightweight Java class, the system is designed for easy integration into software projects and operates without reliance on dictionaries, favoring flexibility and low computational cost. As with other rule-based methods, some limitations remain, particularly with irregular or less frequent word patterns; however, the algorithm is designed to evolve, allowing new rules to be incorporated as issues are identified. Distributed under the GNU GPLv3 license, the library represents a practical and extensible continuation of earlier research on Portuguese syllabification.

3. The Common Orthographic Vocabulary of the Portuguese Language (VOC) and its Brazilian Section VOLP

The Common Orthographic Vocabulary of the Portuguese Language (VOC) is the platform that hosts the instruments legally defining Portuguese orthography. The VOC has been developed under the coordination of the International Portuguese Language Institute (IILP), the CPLP body responsible for linguistic policy, as mandated by the 2010 Brasília Action Plan for the Promotion, Dissemination, and Projection of the Portuguese Language. Established under the 1990 Orthographic Agreement of the Portuguese Language, signed by CPLP countries, the VOC is being progressively implemented across member states. It gradually integrates the vocabulary of each CPLP country, following political validation and adherence to common methodology and criteria, with representative data from Brazil, Cape Verde, Mozambique, Portugal, and Timor-Leste incorporated to date.

The Brazilian component of the VOC corresponds to the *Vocabulário Ortográfico da Língua Portuguesa*[1], with orthographic information specifically revised and adapted. Adaptations for this specific integration included the proper validation of all listed words and the addition of new information, such as syllabic division, stress marking, inflectional paradigms, functional relationships with other entries (specifically morphologically related words), and frequency indices [Bechara 2017]. This adaptation of VOLP for VOC, was the reference adopted in this study to develop Vidya Syl. This Brazilian section [Bechara 2017], specially developed for the VOC, will henceforth be referred to as VOLP-VOC.

Its database is freely and universally available to citizens, accessible via the internet, containing more than 300,000 entries and over two million orthographic forms. It has been developed in accordance with the latest advances in lexicography and consists of a tagged vocabulary that allows users to identify, among other things, the attested usage of each word in each country, making it possible to determine which words are common to all countries and which occur in specific countries [International Institute of the Portuguese Language 2024].

[1]Orthographic Vocabulary of the Portuguese Language. Our translation.

4. Vidya Syl

The Vidya Syl algorithm was developed as an evolution of rule-based approaches for Portuguese syllabification, building primarily on the work of Amorim (2024), which in turn is grounded in the studies of Silva, Braga, and Resende Jr. (2008) and Rocha and Neto (2013). The main goal of Vidya Syl is to advance this line of research by revising and expanding linguistic rules to achieve closer adherence to the VOLP-VOC, ensuring greater consistency with current orthographic standards.

Its implementation was structured in Java 8 with a focus on modularity and efficiency, while preserving independence from dictionary-based resources. One of its key distinguishing features is a right-to-left scanning strategy, which first identifies the vocalic nucleus of each syllable and then determines its surrounding elements, such as adjacent consonants and consonant clusters. This approach enables a segmentation process more closely aligned with the phonological structure of Portuguese, particularly in more complex cases.

The algorithm incorporates refined rules to better handle the patterns codified in VOLP-VOC. As a result, Vidya Syl reduces inconsistencies observed in earlier approaches and improves the accuracy of syllable segmentation in real-world contexts, while preserving the advantages of rule-based methods, such as low computational cost and strong generalization capacity.

4.1. Vidya Syl Architecture

Vidya Syl implements syllabification for Brazilian Portuguese using an algorithm based on successive reduction of the word fragment, applying a decision table grounded in phonetic pattern types as proposed by Amorim (2024), with fallback to the first occurrence of certain transitions (vowel–consonant, consonant–vowel, etc.) and handling of special cases [Amorim 2024, Grandi 2026]. It operates on words or texts, returning syllables separated by hyphens. The core logic relies on an extensive set of rules that analyze the phonetic structure of the word. Given a word, after normalization and removal of punctuation marks, hyphens, and special characters, the algorithm converts the word to lowercase. Then, it generates a phonetic pattern based on Amorim (2024), where each letter is mapped to a category:

- v = vowel (a, e, o, á, é, í, ó, ú, ä, ë, ï, ö, ü)
- s = semivowel (i, u)
- t = tilde vowel (ã, ẽ, ĩ, õ, ũ)
- n = nasal consonant (m, n when not followed by h)
- m = mute consonant (b, c, d, f, g, p, t not followed by l, r, t, s, v)
- o = occlusive consonant (b, c, d, p, g, t followed by l, r, t, s, v; also q)
- l = liquid consonant (r, l when not followed by h)
- d = identical double consonant (rr, ss)
- h = digraph with h (lh, nh)
- c = any other consonant

This pattern is used to locate syllable boundaries by applying a series of specific conditions that handle combinations of up to three or four letters at the beginning of the remaining fragment. If no special rule applies, the algorithm searches for the leftmost position of certain sequences (e.g., “cn”, “co”, “sc”, “vc”, etc.).

The core of the algorithm lies in the method `Syllables(String[] words, boolean verbose)`, which iterates over words and, for each one, repeatedly processes a fragment (initially the entire word) until it becomes empty. At each iteration:

- It obtains the pattern of the current fragment.
- It applies a set of rule-based conditions.
- If no condition is satisfied, it attempts to find the lowest index among several substrings (e.g., "cn", "co", "sc", "vc", "lsc", "nsl", etc.) and splits the fragment accordingly; otherwise, it adds the entire fragment as a syllable.

The algorithm includes hundreds of specific rules, organized into methods named after combinations of phonetic patterns. Each method analyzes the pattern to determine how many letters should be separated as a syllable. The separation is performed by removing the initial characters of the fragment and adding them to the syllable list.

A set of special cases is explicitly addressed, including:

- Non-divisible words: a predefined set of words, initialisms, and acronyms, such as "IPVA", "jeans", and *"yin"*, that are not split.
- Roman numerals: they are preserved as whole units.
- Vowel constraints: if the fragment does not contain at least two vowels, the word is considered indivisible (with exceptions such as "brunch" and "mach").

The algorithm is highly heuristic, covering exceptions and loanwords (e.g., shock, week, brunch), and prioritizes specific rules over the general principle of splitting before a consonant that begins a non-permitted cluster. Its complexity is justified by the need to ensure adherence to the syllabification rules established by the VOLP-VOC.

5. Methodology

This section describes the methodological procedures adopted to develop and evaluate the Vidya Syl algorithm. It includes the large-scale extraction of syllabified entries from the VOLP-VOC, followed by data cleaning steps such as the removal of polylexical lemmas and duplicate records. Subsequently, an initial evaluation was conducted using the baseline algorithm proposed by Amorim (2024), and an iterative process of rule refinement was carried out to maximize alignment with the VOLP-VOC. The section concludes with the criteria used to determine the final level of adherence achieved by the proposed method.

5.1. Retrieval of syllables from the VOLP-VOC

The Vidya Syl class `MineVolpEntries.java` was employed to retrieve all syllabifications available in the VOLP-VOC, which are accessed through lemma queries on the web via Uniform Resource Identifiers (URIs), as illustrated in Figure 1. Given that the highest lemma reference in the VOLP-VOC is 241,289, it systematically queried the entire range of lemma identifiers, from 1 to 241,289.

In this process, 224.238 entries were successfully returned; that is, the search and identification algorithm for the syllables of 17.051 lemmas generated some computational exceptions that could not be processed. For instance, when `n = 1`,

VOLP-VOC returns the content displayed in Figure 2: the verb *cultivar* segmented into the syllables by points.



Figure 1. URI for lemma queries in the VOLP-VOC.

The next step consisted of converting the VOLP-VOC notation of syllabification and lexical stress into the hyphenated form adopted by Vidya Syl. This transformation was performed by the class `AdjustVolpLemmas.java`. For instance, the notation `(cul.ti.'var)` was converted to `cul-ti-var`.



Figure 2. Return of lemma 1 from the VOLP-VOC.

5.2. Removal of Multiword, Unattested, and Duplicate Lemmas

As discussed by Nunes (2022), polylexicality occurs when a lexical unit (defined as a word or a sequence of words with a single meaning that is treated as an element of a language's lexicon) is composed of at least two lexical elements (words). The query to the VOLP-VOC lemmas returned 43 polylexical lemmas, which were removed from the lemmas set [Nunes 2022].

For 19 words, the VOLP-VOC provides different syllabifications depending on the grammatical class. These words are: *adônio, afidídeo, aguaiano, ajuizador, antidiurético, atribuidor, banyamulenge, braille, carajá-mirim, chauffeur, haeckeliano, iao, insetófilo, junkie, macauense, picuipeba, saudar, wronskiano,* and *zombie*. In such cases, one syllabification per word was selected at random. From this process, it was removed 24 entries.

Following manual revision via the VOLP-VOC query interface, five additional entries flagged as not attested were removed. These words are: *fröbeliano, ci-iao, fecho-éclair, hors-d´oeu-vre,* and *codo-d´áua*.

Many words were registered under more than one grammatical class while sharing the same syllabification. The word *tal*, for instance, can function in several distinct grammatical classes, as shown in Figure 3. After disregarding duplicate entries corresponding to words that share the same syllabification across distinct grammatical classes, 186,241 unique words and their respective syllabifications remained.

5.3. Initial Evaluation at Amorim's Code

Before initiating the process of incorporating phonological and orthographic rules to align Vidya Syl with the VOLP-VOC, an evaluation was conducted on the selected



Figure 3. Different classes of word *tal* in the VOLP-VOC.

words and their corresponding syllabifications using Amorim's original algorithm. In this initial phase, it was found that 49,467 (26.56%) of the 186,241 available syllabic divisions were divergent.

5.4. Incremental Insertion of Morphological and Ortographic Constraints in Vidya Syl

An extensive process of incorporating phonological and orthographic rules into Vidya Syl was then initiated. After each new rule was added, the alignment rate with the VOLP-VOC was recomputed to determine whether it had increased or decreased. All 186,241 entries were verified by the Vidya Syl class `CheckSyllabesVolp.java`, with each iteration taking approximately 13 seconds. When alignment improved, the rule was retained; when it decreased, the rule was discarded.

5.5. Finalization

Figure 4 shows the final iteration of improvements, in which 186,186 entries were marked as `ok` (i.e., compliant with the VOLP-VOC) and 55 entries as `wrong` (i.e., non-compliant with the VOLP-VOC).

```
partial: 00:00:01       total:10000     ok: 10000       wrong:0
partial: 00:00:02       total:20000     ok: 19998       wrong:2
partial: 00:00:02       total:30000     ok: 29998       wrong:2
partial: 00:00:03       total:40000     ok: 39997       wrong:3
partial: 00:00:04       total:50000     ok: 49995       wrong:5
partial: 00:00:05       total:60000     ok: 59995       wrong:5
partial: 00:00:05       total:70000     ok: 69993       wrong:7
partial: 00:00:06       total:80000     ok: 79993       wrong:7
partial: 00:00:07       total:90000     ok: 89990       wrong:10
partial: 00:00:07       total:100000    ok: 99987       wrong:13
partial: 00:00:08       total:110000    ok: 109986      wrong:14
partial: 00:00:09       total:120000    ok: 119983      wrong:17
partial: 00:00:10       total:130000    ok: 129982      wrong:18
partial: 00:00:10       total:140000    ok: 139980      wrong:20
partial: 00:00:11       total:150000    ok: 149977      wrong:23
partial: 00:00:11       total:160000    ok: 159965      wrong:35
partial: 00:00:12       total:170000    ok: 169954      wrong:46
partial: 00:00:13       total:180000    ok: 179952      wrong:48
VOLP
final.: 00:00:13        total:186241    ok: 186186      wrong:55
```

Figure 4. URI for lemma queries in the VOLP-VOC.

Considering the 186,241 syllabifications retrieved from the VOLP-VOC, the alignment process was concluded when Vidya Syl reached 99.97% agreement with the reference (186,191 out of 186,241 entries). The remaining 55 words could not be aligned due to uncertainties regarding their correct syllabic representation in the VOLP-VOC itself. The verification of these 55 uncertain cases was deferred to future work, to be carried out jointly with linguists specialized in phonology and lexicography:

1. Table 1 lists six words for which doubts remain regarding the separation of double consonants;

2. Table 2 lists eleven words for which uncertainties persist regarding the syllabic segmentation of the falling diphthong *au*;
3. Table 3 lists four words for which uncertainties persist regarding the syllabic segmentation of hiatuses;
4. Table 4 lists nine words with doubts regarding Portuguese orthographic rules;
5. Table 5 lists three words for which uncertainties persist regarding the syllabification of English loanwords; and
6. Table 6 lists 22 words whose syllabic division conflicts with similar ones in the VOLP-VOC.

Table 1. Words with doubts regarding the separation of double consonants.

Word	VOLP-VOC Syllables	Word	VOLP-VOC Syllables
antisséptico	an-tissép-ti-co	rickettsiácea	ric-kett-si-á-cea
rickettsíase	ric-kett-sí-a-se	rickettsiáceo	ric-kett-si-á-ceo
rickettsiose	ric-kett-si-o-se	wölffliniano	wölffli-ni-a-no

Table 2. Words with doubts regarding the segmentation of the falling diphthong *au*.

Word	VOLP-VOC Syllables	Word	VOLP-VOC Syllables
ressaudação	res-sa-u-da-ção	saudador	sa-u-da-dor
ressaudado	res-sa-u-da-do	saudável	sa-u-dá-vel
ressaudar	res-sa-u-dar	saudosismo	sa-u-do-sis-mo
saudação	sa-u-da-ção	saudosista	sa-u-do-sis-ta
saudade	sa-u-da-de	saudoso	sa-u-do-so
saudado	sa-u-da-do		

Table 3. Words with doubts regarding the separation of double consonants.

Word	VOLP-VOC Syllables	Word	VOLP-VOC Syllables
coesposo	coes-po-so	reengineering	reen-gi-nee-ring
sacroilíaco	sa-croi-lí-a-co	yeti	ye-ti

Table 4. Words with doubts regarding Portuguese orthographic rules.

Word	VOLP-VOC Syllables	Word	VOLP-VOC Syllables
araquãí	a-ra-qu-ã-í	reganho	re-gan-ho
icozeiro	i-cozei-ro	tubinar	tu-bi-na-r
malcoado	malco-a-do	tupperware	tup-pe-rwa-re
quão	qu-ão	westphalense	westp-ha-len-se
reaquisto	re-a-qui-sto		

Table 6. Words involving uncertainty regarding similar terms in the VOLP-VOC.

Word	VOLP-VOC Syllables	VOLP-VOC Similar
business	busi-ness	`("co.re-bu.'si.ness)`
contraemergentes	con-trae-mer-gen-tes	`(con.tra.e.mer.*gen.te)`
diretora-geral	di-re-tora-ge-ral	`(di.re.'to.ra)`
email	e-ma-il	`("e-'mail)`
físico-quimicamente	fí-si-co-qu-imi-ca-men-te	`(qui.mi.ca.'men.te)`
funboard	fun-bo-ard	`(clip;'board)`
geográfico-linguisticamente	ge-o-grá-fi-co-lin-gu-isti-ca-men-te	`(lin.guis.ti.ca.'men.te)`
heautognóstico	he-au-tog-nós-ti-co	`(heau;to.'gno.se)`
keynesiano	ke-y-ne-si-a-no	`(key;ne.si.a.'nis.mo)`
magnificat	ma-gni-fi-cat	`(mag.ni.fi.'ca.do)`
miudinha	miu-di-nha	`(mi.u.*di.nho)`
neomalthusianista	ne-o-malt-hu-si-a-nis-ta	`(mal;thu.si.'a.no)`
pipuiruçu	pi-pui-ru-çu	`(pe.pu.i.ru.*çu)`

Word	VOLP-VOC Syllables	VOLP-VOC Similar
pituitária	pi-tui-tá-ria	(pi.tu.i.*tá.ri,o)
poliuretana	po-liu-re-ta-na	(po.li.u.re.*ta.no)
rato-da-índia	ra-to-da-ín-di-a	('ín.di,a)
sã-luisense	sã-lu-i-sen-se	(lui.*sen.se)
são-luisense	são-lu-i-sen-se	(lui.*sen.se)
serial-killer	se-rial-kil-ler	(se.ri.'al)
suicidomaníaco	sui-ci-do-ma-ní-a-co	(su.i.*ci.da)
tapuirana	ta-pui-ra-na	(a.pu.i.*ra.na)
wesleyanismo	wes-ley-a-nis-mo	(wes.le.y.a.*nis.ta)

6. Results

The 99.97% alignment achieved with the VOLP-VOC was considered a final point for the first version of Vidya Syl. When consulting linguistics and phonology specialists from the Linguistics Department of UFRGS, they were positive about the work, but had difficulty confirming the accuracy of the syllabic divisions provided by the VOLP-VOC for the 55 uncertain words. This validation may be revisited at a later date.

6.1. Impact on Flesch Reading Ease (FRE) Index

The Flesch Reading Ease (FRE) is a widely used readability metric that evaluates textual comprehensibility based on average sentence length and average word length (in syllables). It yields scores from 0 to 100, where higher values indicate easier texts and lower values denote greater complexity. Originally proposed by Rudolf Flesch, it is broadly applied in education and natural language processing [Flesch 1948].

The study conducted by Martins et al. (1996) focuses on the adaptation of the FRE Index to Brazilian Portuguese, translating and adjusting the original formula to reflect the specific linguistic characteristics of the language, and proposing the following adapted equation [Martins et al. 1996]:

$$\textbf{Index} = 248.835 - 1.015 \times \left(\frac{\text{words}}{\text{sentences}}\right) - 84.6 \times \left(\frac{\text{syllables}}{\text{words}}\right) \quad (1)$$

The Portuguese text analyzed by the tools in this evaluation:

> *Foi o senador Flávio Arns (PT-PR) quem sugeriu a inclusão da peça entre os itens do uniforme de alunos dos ensinos Fundamental e Médio nas escolas municipais, estaduais e federais. Ele defende a medida como forma de proteger crianças e adolescentes dos males provocados pelo excesso de exposição aos raios solares. Se a ideia for aprovada, os estudantes receberão dois conjuntos anuais, completados por calçado, meias, calça e camiseta.*

Using the equation proposed by Martins et al. (1996), along with a dedicated syllabification algorithm, NILC Metrix NLP Tool [Venturus 2024] found 3 sentences, 69 words, 23 words by sentence and 2,31 syllables by word. Using these parameters, the resulting FRE Index score was **29.3**.

Using the same equation proposed by Martins et al. (1996), Vidya Text [Grandi et al. 2026], an Educational Tool that employs Vidya Syl as its syllabification algorithm, found 3 sentences, 68 words, 22,7 words by sentence and 2,3 syllables by word. Using these parameters, the resulting FRE Index score was **31.7.**

The different results obtained with the same equation, using different syllabification algorithms, highlight the impact of this factor on the computation of readability indices.

Table 5. Words with doubts regarding the syllabification of English loanwords.

Word	VOLP-VOC Syllables	Word	VOLP-VOC Syllables
washingtónia	was-hing-tó-nia	washingtônia	was-hing-tô-nia
washingtonite	was-hing-to-ni-te		

7. Conclusions and Future Works

This study presented Vidya Syl, a syllable division algorithm for Brazilian Portuguese, designed to achieve high adherence to the VOLP-VOC. Building upon prior rule-based approaches, particularly the algorithm proposed by Amorim, the proposed method introduced a hybrid strategy that combines phonological and orthographic rules with large-scale lexical validation. This combination proved effective in addressing known limitations of purely rule-based systems, especially with regard to exception handling and normative consistency.

The results demonstrate that Vidya Syl achieves a high level of alignment with the VOLP-VOC, reaching 99.97% agreement over a dataset of 186,241 lexical entries. This represents a significant improvement over the baseline algorithm, which exhibited 26.56% divergence in the initial evaluation. The incremental rule insertion strategy, guided by empirical validation against the VOLP-VOC, was instrumental in achieving this level of performance, while preserving the advantages of rule-based methods, such as low computational cost, interpretability, and independence from large lexical databases at runtime.

Despite these advances, the analysis also revealed inherent challenges in the standardization of syllabification. A small subset of 55 words could not be conclusively aligned due to uncertainties in their syllabic representation within the VOLP-VOC itself, even when evaluated by domain experts. This finding highlights not only the complexity of syllabification in Portuguese but also potential limitations or ambiguities in normative lexical resources.

Future work will focus on many directions: i) to rely on expert judgment to settle the 55 words whose syllabification remained inconclusive; ii) efforts will be made to optimize the algorithm; and iii) integrate it into NLP applications, such as readability tools and educational technologies. Its code will be made available for these purposes and also to enable other research groups to carry out similar work with other Portuguese dialects in relation to the VOLP-VOC.

In summary, Vidya Syl represents a robust and extensible contribution to Portuguese NLP, demonstrating that hybrid approaches grounded in both linguistic theory and lexical validation can achieve high levels of accuracy and practical relevance. Furthermore, its results already feed into a first version of the Vidya Text Tool, which indicates the readability potential of a text according to the Brazilian Flesch index formula, but now with a new index in the syllable count.

8. Acknowledgments

This research was supported by CNPq (National Council for Scientific and Technological Development), grant PQ 307088/2023-5, National Institutes of Science and Technology – CNPq-TILD-IAR (grant 408490/2024-1), PROBIC FAPERGS-

UFRGS, and FAPERGS Ed. 06/2025 - PROGRAMA PESQUISADOR GAÚCHO – PqG (grant 25/2551-0002702)

9. Ethics Statement

No sensitive personal data were used in the development of the Vidya Syl algorithm. The research is aligned with the ethical principles of the field of Computational Linguistics and aims to provide the community with a properly tested syllabification algorithm.

10. Limitations

A limitation encountered during this study is that VOLP-VOC occasionally assigns different syllabifications to the same lexical item according to its grammatical category. For example, braille is represented as a masculine noun (’brail;le) and as an adjective (bra.il.le). A further limitation concerns the verification of correct syllabification across the full inventory of Brazilian Portuguese lexical entries. Despite this challenge, the proposed algorithm achieved full agreement with the syllabifications provided in the VOLP-VOC, except for 55 words whose correct syllabification remains uncertain.

11. Statement on the Use of Generative AI

In accordance with CNPq Ordinance No. 2,664/2026, the use of the GPT-5.5 model of ChatGPT is hereby declared as a support tool for stylistic refinement and for translating the original text from Portuguese into English. The authors are fully responsible for the final content presented.

References

Amorim, E. (2024). syllable-pt: Portuguese syllabification algorithm [source code]. `https://github.com/evelinamorim/syllable-pt`.

Bechara, E., editor (2017). *Vocabulário Ortográfico da Língua Portuguesa*. Academia Brasileira de Letras, Rio de Janeiro, 5 edition. Edição adaptada ao VOC.

Câmara Jr., J. M. (1992). *Estrutura da língua portuguesa*. Vozes.

Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Grandi, R. H. (2026). Vidya syl: A syllable division algorithm adhering to the brazilian section of the common orthographic vocabulary of the portuguese language (volp). `https://gitlab.com/rogesgrandi/vidyasyl`. Accessed: 19 Jul. 2026.

Grandi, R. H. et al. (2026). Vidya text: an educational tool to support the teaching of writing and the study of brazilian portuguese vocabulary. https://vidyatext.nuvem.ufrgs.br. Acesso em: 19 jul. 2026.

International Institute of the Portuguese Language (2024). Voc: Vocabulário ortográfico comum da língua portuguesa. `https://voc.cplp.org/index.php?action=von&von=br`.

International Institute of the Portuguese Language (2026). Vocabulário ortográfico comum da língua portuguesa. https://voc.cplp.org/index.php?action=ficha. Comunidade dos Países de Língua Portuguesa (CPLP).

Martins, T. B., Ghiraldelo, C. M., Nunes, M. d. G. V., and Oliveira Junior, O. N. d. (1996). Readability formulas applied to textbooks in brazilian portuguese.

Nunes, Q. O. B. (2022). *Unidades fraseológicas em Sagarana de Guimarães Rosa: Um estudo de traços culturais em tradução ao francês*. Tese de doutorado, Universidade Federal de Mato Grosso do Sul.

Rocha, W. and Neto, N. (2013). Implementação de um separador silábico gratuito baseado em regras linguísticas para o português brasileiro. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology (STIL)*.

Silva, D. C., Braga, D., and Jr., F. G. V. R. (2008). Separação das sílabas e determinação da tonicidade no português brasileiro. In *XXVI Simpósio Brasileiro de Telecomunicações (SBrT)*, pages 1–5.

Venturus (2024). Nilc-metrix. https://pln.venturus.dev.br/nilcmetrix. Accessed: 6 abr. 2026.

Developing Multilingual Multiword Expressions (MWEs) Resources: An Annotated Corpus of Fables following the PARSEME guidelines

Letícia Guedes Guimarães[1], Adriana Pagano[1], André V. Lopes Coneglian[1], Aline Villavicencio[2]

[1] Universidade Federal de Minas Gerais
[2]University of Exeter
{letigg, apagano, coneglian}@ufmg.br
a.villavicencio@exeter.ac.uk

Abstract. *This paper introduces a Brazilian Portuguese corpus of fables developed as part of a broader multilingual initiative. It comprises the narratives and their concluding moral statements (epimythia), annotated for morphosyntax and multiword expressions (MWEs) according to the Universal Dependencies (UD) and PARSEME guidelines, respectively. The analysis reveals that 53.59% of the sentences contain at least one MWE, with inherently reflexive verbs (IRVs) constituting the most frequent category. MWEs are also distributed differently: narratives contain a higher proportion of adverbial MWEs, whereas adpositional MWEs are frequent in epimythia. Future work will enrich the corpus through semantic annotation based on Uniform Meaning Representation (UMR).*

1. Introduction

Multiword expressions (MWEs) constitute a persistent challenge in Natural Language Processing (NLP) due to their idiosyncratic grammatical properties [Savary et al. 2018][Pasquer et al. 2020]. Their accurate identification is crucial for tasks such as parsing, machine translation, and semantic analysis, as failure to capture these expressions often leads to systematic errors in meaning interpretation.

Despite their importance in multilingual NLP resource development, corpora annotated with MWEs are limited across languages and text types [Ramisch et al. 2026]. Addressing this limitation requires text-type-diverse corpora; in this context, fables are particularly suitable for corpus construction because they are compact, self-contained discourse units that combine narrative discourse, dialogue, and reported speech with figurative and evaluative moral statements, thus moving from particular events to generalized meanings. This structure makes it possible to relate MWE use to different narrative stages and communicative functions. Moreover, fables circulate across languages and cultures, often in multiple adaptations, supporting multilingual and contrastive analysis.

This paper introduces an annotated corpus of Aesop's fables translated into Brazilian Portuguese as part of a larger multilingual corpus of fables translated into French, English, Spanish, Russian, Chinese, and Italian. The corpus comprises texts translated from Greek and annotated for morphosyntactic categories using the Universal Dependencies (UD) framework [Nivre et al. 2020] and for MWEs following the PARSEME guidelines [Savary et al. 2017] [Savary et al. 2026]. It is designed to support multilingual NLP analysis.

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 200–208
October 19-22, 2026 ©2026 Association for Computational Linguistics

2. Related work

Annotated corpora enriched with linguistic information are essential for both training and evaluating Natural Language Processing (NLP) models [Ramisch et al. 2026]. A foundational initiative in this domain is the UD framework [Nivre et al. 2020], which provides standardized morphosyntactic annotation across languages, enabling consistent training and evaluation of multilingual parsers and treebanks. Building on UD, the PARSEME initiative [Savary et al. 2017] introduced guidelines for the annotation of MWEs, initially focusing on verbal MWEs and later extending to other categories, including nominal and adjectival expressions [Scholivet et al. 2026]. This effort resulted in gold-standard corpora that are widely used to evaluate MWE identification systems across multiple languages [Savary et al. 2026]. Details of the Brazilian Portuguese annotation process in Parseme versions 1.0 and 2.0, including challenging annotation cases, are provided in [Ramisch et al. 2018] and [Savary et al. 2026]; annotation has so far focused primarily on news reports, which highlights the need for annotated corpora covering other text types, which we addtress by annotating fables.

3. Methodology

For this multilingual corpus, the Brazilian Portuguese translations used are those published in *Esopo: fábulas completas* [Dezotti et al. 2013]. As the fables were available only in print, the first step consisted in manually transcribing them in a text editor. The texts were saved as plain-text (.txt) files and organized in two subsets: narratives and morals (epimythia). Overall, the corpus comprises 418 sentences and 8443 tokens excluding punctuation. The distribution of both subsets can be found in Table 1.

Table 1. Corpus statistics

Corpus	Sentences	Tokens	Types	Lemmas	Avg. sentence length
Narratives	355	7196	2152	1668	20.27
Epimythia	63	1247	507	438	19.79
Full corpus	418	8443	2427	1843	20.2

Subsequently, all .txt files were automatically parsed using UDPipe [Straka et al. 2016], trained on the Petrogold corpus [Souza et al. 2021]. The resulting annotated CONLLU files were then uploaded to the platform Arborator Grew [Guibon et al. 2020] and manually revised following the Brazilian Portuguese guidelines for UD annotation [Duran 2021, Duran 2022].

The revised files were uploaded to FLAT (FoLiA Linguistic Annotation Tool) [van Gompel 2024] for manual annotation of MWEs. Following the PARSEME framework [Savary et al. 2026], the annotation process included all MWE categories established by the project, which encompass verbal, nominal, adjectival, adverbial, and functional expressions[1]. Table 2 presents all MWE categories identified in the corpus, together with illustrative examples. The full set of categories, along with def-

[1]In compliance with PARSEME guidelines for Brazilian Portuguese [Savary et al. 2026], idiomatic verb-particle constructions (IVPCs) were not annotated. The guidelines report only one such construction in the language, namely "jogar fora" (lit. 'to throw outside'; idiomatically, 'to throw away' or 'to discard'), which was reclassified under a different label (VID) and the IVPC category excluded.

initions, identification tests, and examples, can be found in the PARSEME repository [PARSEME Initiative 2026].

Table 2. PARSEME categories adopted in this study

Type	Label	Name	Example and gloss
Verbal	IRV	Inherent reflexive verb	*se queixar* 'to complain'
	LVC.full	Light-verb construction (full)	*fazer uma visita* (lit. 'to make a visit'; 'to pay a visit')
	LVC.cause	Light-verb construction (causative)	*dar preferência* (lit. 'to give preference'; 'to prefer')
	VID	Verbal idiom	*correr riscos* (lit. 'to run risks'; 'to take risks')
	IAV	Inherently adpositional verb	*tratar de* (lit. 'to treat of'; 'to deal with')
	MVC	Multi-verb construction	*quer dizer* (lit. 'to want to say'; 'to mean')
Nominal	NID	Nominal idiom	*cara de pau* (lit. 'wooden face'; 'shameless person')
Adjectival / Adverbial	AdjID	Adjectival idiom	*cheia de si* (lit. 'full of herself'; 'conceited')
	AdvID	Adverbial idiom	*de novo* (lit. 'of new'; 'again')
Functional	AdpID	Adposition idiom	*a partir de* 'from; starting from'
	ConjID	Conjunction idiom	*uma vez que* 'since; given that'
	IntjID	Interjection idiom	*pois é* 'indeed; well, yes'

The annotation was carried out by the first author - a PhD candidate in Linguistics - and subsequently revised by the co-authors, senior researchers in Linguistics and Computational Linguistics. A blind procedure was not adopted due to the high level of expertise demanded by the annotation task. Disagreements were resolved through discussion guided by the PARSEME guidelines [Savary et al. 2026]. The resulting revised XML files were converted to XLSX format using a customized Python script and stored in a dedicated folder.

Quantitative analysis was conducted using two customized Python scripts. The first processed the CoNLL-U narrative and epimythia files into CSV format to extract structural metrics, including token, type, and lemma counts, as well as average sentence length (Table 1). The second processed the XLSX files to compute absolute and relative MWE frequencies, average MWEs per sentence, sentence-level MWE occurrence, and average MWE length. A lemmatization dictionary was used to map surface variants to canonical forms - for example, the forms "se pôs" (REFL put.PST.PFV.3SG, 'began') and "se punha" (REFL put.PST.IPFV.3SG, 'was beginning') were lemmatized as "pôr-se" (put.INF-REFL, 'to begin'). The resulting XLSX files were also used for exploratory qualitative analysis.

4. Results and Discussion

4.1. MWE density and distribution

The corpus comprises 322 MWE occurrences, with 245 distinct canonical forms, as established through the lemmatization procedure described in the Methodology; on average, each sentence contains 0.77 MWEs. Overall, 224 sentences contain at least one annotated MWE, corresponding to 53.59% of the dataset (Table 3). When narratives and epimythia are considered separately, the relative frequency of MWE occurrences remains similar in both, indicating a consistent presence of multiword phenomena across the corpus.

Table 3. Distribution of MWEs across the corpus

	MWE occurrences			Sentences with MWE		
	Total	**Distinct**	**Avg. MWE length**	**Total**	**Rel. freq. (%)**	**Avg. MWEs per sentence**
Narratives	268	206	2.35	189	53.24	0.75
Epimythia	54	47	2.33	35	55.56	0.86
Full corpus	322	245	2.34	224	53.59	0.77

The average length of MWEs in the corpus is 2.34 tokens, with the longest instances spanning 5 tokens: the adverbial idioms (AdvID) "A bem de a verdade" (lit. 'for the good of the truth'; idiomatically, 'truth be told') and "Em o que respeita a" (lit. 'in what respects to'; idiomatically,'regarding'). A total of 72% of occurrences span two tokens, a pattern that is heavily influenced by category distribution, as inherently reflexive verbs (IRVs) and light-verb constructions (LVCs) account for a considerable part of all MWE occurrences. IRVs are strictly formed by two lexicalized (fixed) items: a verb + a clitic (e.g., "tornar-se" ('to become'). As for LVCs, their structure is more flexible, since they can include a determiner or a preposition between the semantically bleached verb and the abstract predicative noun, causing them to span up to 4 tokens in Brazilian Portuguese. However, in our corpus, 70% of the LVC occurrences span only 2 tokens, e.g., "ter ideias" (lit. 'to possess ideas'; idiomatically, 'to come up with ideas').

4.2. MWE category distribution

IRVs account for the majority of MWE occurrences in the corpus, totaling 102 occurrences, with 73 distinct MWE lemmas. Overall, they account for 31.68% of all MWE instances. This MWE category is seconded by LVC.full (16.46%). The total number of occurrences for each categoryis presented in Table 4.

The distribution of MWE categories varies between the two fable sections. As the comparative Table 5 shows, while the narratives feature all twelve categories identified in the corpus, the epimythia lack any instances of AdjIDs, IAVs, or MVCs. It is worth noting, however, that these three categories are also scarce within the narrative data. Accounting for the scarcity of AdjIDs is beyond the scope of this paper, but the low frequency of IAVs and MVCs can be partially explained by both the annotation guidelines and linguistic factors.

In the case of IAVs, the primary cause likely stems from the annotation guidelines. This category covers verbs that require an adposition to exist or to hold a specific

Table 4. Distribution of MWE categories in the full corpus

	Occurrences			
Category	**Total**	**Rel. freq. (%)**	**Distinct**	**Rel. freq. (%)**
IRV	102	31.68	73	29.67
LVC.full	53	16.46	50	20.33
AdvID	48	14.91	35	14.23
AdpID	46	14.29	26	10.57
VID	30	9.32	27	11
IAV	12	3.73	7	2.85
ConjID	9	2.80	7	2.85
AdjID	8	2.48	8	3.25
NID	5	1.55	5	2.03
IntjID	5	1.55	4	1.63
LVC.cause	3	0.93	3	1.22
MVC	1	0.31	1	0.41

meaning, such as "tratar de" (lit.'to treat of'; idiomatically 'be about'). However, at the current stage of the project, this category remains experimental, and no language-specific guidelines have been established yet to systematically account for its occurrences. Consequently, the seven identified instances probably do not represent the full range of IAV constructions in the corpus, suggesting that their low frequency may partially reflect underannotation. As for MVCs, such as "querer dizer" (lit.'to want to say'; idiomatically, 'to mean'), there are only two occurrences of this category in the Brazilian Portuguese portion of the PARSEME corpus [Savary et al. 2026], suggesting that these constructions tend to be rare. The observed distribution in our corpus is consistent with this tendency.

Regarding the two remaining non-functional MWE categories, VIDs and NIDs, they account for only 10% of all occurrences, proving fairly rare in the corpus. Similarly to the case of AdjIDs, accounting for this scarcity is beyond the scope of this article; hence, a detailed investigation remains an upcoming step in our research.

4.3. Comparing narrative and epimythia: AdvIDs and AdpIDs

Although IRVs predominate across both fable sections, AdvIDs rank second in the narratives, whereas AdpIDs take second place in the epimythia (Table 5). This distribution directly reflects the distinct grammatical configurations characterizing these two functional sections [Neves 2014].

While narratives are primarily concerned with the unfolding of events, epimythia focus on expressing the moral lesson derived from them. Accordingly, narratives tend to favor adverbial idioms, which contribute to event sequencing, as in *em seguida* (lit. 'in sequence'; idiomatically, 'next' or 'then'); locate events in time and space, as in *tempos depois* (lit. 'times later'; idiomatically, 'some time later'); and express manner and situational circumstances, as in *em silêncio* (lit. 'in silence'; idiomatically, 'silently') and *de presente* (lit. 'of gift'; idiomatically, 'as a gift').

In contrast, epimythia show a stronger preference for adpositional idioms, which

encode relations of reference, as in *em relação a* (lit. 'in relation to'; idiomatically, 'with regard to'); opposition, as in *ao contrário de* (lit. 'to the contrary of'; idiomatically, 'in contrast to'); cause, as in *devido a* (lit. 'due to'; idiomatically, 'because of'); and purpose, as in *em busca de* (lit. 'in search of'; idiomatically, 'seeking'). These relations play a central role in the formulation of arguments and generalizations. This interpretation is further supported by the near absence of AdvIDs in epimythia, where the only occurrence is *às claras* (lit. 'at the clear [things]'; idiomatically, 'openly' or 'in plain sight'). In narratives, by contrast, AdpIDs primarily provide spatial, causal, and instrumental framing, as in *ao pé de* (lit. 'at the foot of'; idiomatically, 'next to'), *diante de* (lit. 'in front of'; idiomatically, 'faced with'), and *por meio de* (lit. 'by means of'; idiomatically, 'through').

Table 5. Distribution of MWE categories in narratives and epimythia

	Narratives			Epimythia			
Category	**Total**	**%**	**Rank**	**Total**	**%**	**Rank**	**Rank diff. (N−E)**
IRV	84	31.34	1	18	33.33	1	0
AdvID	47	17.54	2	1	1.85	9	-7
LVC.full	46	17.16	3	7	12.96	4	-1
AdpID	35	13.06	4	11	20.37	2	2
VID	21	7.84	5	9	16.67	3	2
IAV	12	4.48	6	0	0.00	—	—
AdjID	8	2.99	7	0	0.00	—	—
ConjID	7	2.61	8	2	3.70	5	3
NID	3	1.12	9	2	3.70	5	4
IntjID	3	1.12	9	2	3.70	5	4
LVC.cause	1	0.37	10	2	3.70	5	5
MVC	1	0.37	10	0	0.00	—	—

5. Conclusion and future work

This paper presented a manually annotated corpus of Aesop's fables translated into Brazilian Portuguese, developed as part of a broader multilingual initiative. The corpus was annotated for UD [Nivre et al. 2020] and for MWEs following the PARSEME guidelines [Savary et al. 2017][Savary et al. 2026], aiming to support multilingual and cross-text type NLP research.

The quantitative analysis revealed that 53.59% of the sentences contain at least one annotated MWE. IRVs account for the majority of instances across both narratives and epimythia. However, when comparing these two fable sections, adpositional MWEs are more frequent in the epimythia, whereas narratives favor adverbial MWEs. This reflects the distinct grammatical configurations of these two sections [Neves 2014].

Future work includes enriching the corpus with semantic annotation following the Universal Meaning Representation (UMR) framework [Chun and Xue 2024] and deepening the qualitative analysis to address open questions from the results and discussion section. The fully annotated dataset and the scripts used for data processing and analysis will be made publicly available in a GitHub repository under CC BY 4.0 and the MIT License, respectively.

6. Acknowledgments

The authors express their gratitude to the anonymous reviewers for their constructive feedback. Letícia Guedes acknowledges the financial support of the Minas Gerais State Agency for Research and Development (FAPEMIG). Adriana Pagano's research is funded by FAPEMIG and the National Council for Scientific and Technological Development (CNPq, grants 404722/2024-5; 313103/2021-6). André V. Lopes Coneglian holds a grant from CNPq (153972/2025-4). This work received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology) and in conducted under the auspices of the National Institute of Science and Technology in Responsible Artificial Intelligence for Computational Linguistics, Information Treatment, and Dissemination (INCT-TILDIAR), funded by CNPq (grant 408490/2024-1).

7. Limitations

The relatively small size of the corpus restricts broader linguistic generalizations.

8. Ethics statement

The study does not involve human participants, personal data, or any type of sensitive information. All materials used consist solely of publicly available linguistic resources, including a book whose translated version was used with the explicit permission of its translator.

9. Generative AI usage

Generative AI tools were used in this work to generate Python code for data analysis and to assist with structuring sections and improving English language expression. All content was subsequently reviewed and validated by the authors. No content or interpretations were generated by AI.

References

Chun, J. and Xue, N. (2024). Uniform meaning representation parsing as a pipelined approach. In Ustalov, D., Gao, Y., Panchenko, A., Tutubalina, E., Nikishina, I., Ramesh, A., Sakhovskiy, A., Usbeck, R., Penn, G., and Valentino, M., editors, *Proceedings of TextGraphs-17: Graph-based Methods for Natural Language Processing*, pages 40–52, Bangkok, Thailand. Association for Computational Linguistics.

Dezotti, M. C. C., Berliner, E., and Duarte, A. (2013). *Esopo: fábulas completas*. Cosac Naify, São Paulo, Brazil.

Duran, M. (2021). Manual de anotação de pos tags: orientações para anotação de etiquetas morfossintáticas em língua portuguesa, segundo as diretrizes da abordagem universal dependencies. Technical Report Relatório Técnico No. 434, Universidade de São Paulo, ICMC. Accessed: 2026-04-10.

Duran, M. (2022). Manual de anotação de relações de dependência: versão revisada e estendida. Technical Report Relatório Técnico No. 440, Universidade de São Paulo, ICMC. Accessed: 2026-04-30.

Guibon, G. et al. (2020). When collaborative treebank curation meets graph grammars. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC)*, pages 5291–5300.

Neves, M. H. M. (2014). A gramática pela fábula. ou: A fábula pela gramática. *Linguística*, 30(1):165–196.

Nivre, J. et al. (2020). Universal dependencies v2: An evergrowing multilingual treebank collection. *arXiv preprint arXiv:2004.10643*.

PARSEME Initiative (2026). Parseme corpora wiki: Citing parseme. `https://gitlab.com/parseme/corpora/-/wikis/home#citing-parseme`. Accessed: 2026-05-19.

Pasquer, C., Savary, A., Ramisch, C., and Antoine, J.-Y. (2020). Verbal multiword expression identification: Do we need a sledgehammer to crack a nut? In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)*, Barcelona, Spain. HAL Id: hal-03013636.

Ramisch, C. et al. (2018). A corpus study of verbal multiword expressions in brazilian portuguese. In *Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018*, Lecture Notes in Artificial Intelligence, Cham, Switzerland. Springer International Publishing.

Ramisch, R., Ramisch, C., and Villavicencio, A. (2026). Expressões multipalavras: fogo de palha ou osso duro de roer? In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, volume 1, book chapter 5. BPLN, 4 edition.

Savary, A. et al. (2017). The parseme shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions*, pages 31–47. Association for Computational Linguistics.

Savary, A. et al. (2018). Parseme multilingual corpus of verbal multiword expressions. In Markantonatou, S., Ramisch, C., Savary, A., and Vincze, V., editors, *Multiword Expressions at Length and in Depth: Extended Papers from the MWE 2017 Workshop*, pages 87–147. Language Science Press, Berlin.

Savary, A., Scholivet, M., Ramisch, C., Nakamura, T., Bilinski, E., Stymne, S., Giouli, V., Markantonatou, S., Pais, V., Mitrofan, M., Estève, L., Guillaume, B., Mititelu, V. B., Čibej, J., Hernández, R. D., Fendel, V., Gantar, P., Kanishcheva, O., Krstev, C., Liebeskind, C., Lobzhanidze, I., Marković, A. M., Nešpore-Bērzkalne, G., Pagano, A. S., Shamsfard, M., Stankovic, R., Tajalli, V., Tiberius, C., and Padhye, A. (2026). Parseme 2.0 multilingual corpus of multiword expressions. In Piperidis, S., Bel, N., van den Heuvel, H., Ide, N., Krek, S., and Toral, A., editors, *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 4819–4834, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Scholivet, M., Savary, A., Ramisch, C., Bilinski, E., Nakamura, T., Mitrofan, M., and Pais, V. (2026). Edition 2.0 of the PARSEME shared task on multilingual identification and paraphrasing of multiword expressions. In Ojha, A. K., Mititelu, V. B., Constant, M., Stoyanova, I., Doğruöz, A. S., and Rademaker, A., editors, *Proceedings of the 22nd*

Workshop on Multiword Expressions (MWE 2026), pages 254–275, Rabat, Marocco. Association for Computational Linguistics.

Souza, E., Silveira, A., Cavalcanti, T., Castro, M., and Freitas, C. (2021). PetroGold – corpus padrão ouro para o dominio do petroleo. In Ruiz, E. E. S. and Torrent, T. T., editors, *Proceedings of the 13th Brazilian Symposium in Information and Human Language Technology*, pages 29–38, Porto Alegre, Brazil. Association for Computational Linguistics.

Straka, M., Hajič, J., and Straková, J. (2016). Udpipe: Trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, pages 4290–4297.

van Gompel, M. (2024). Folia linguistic annotation tool (version 0.11.5). KNAW Humanities Cluster CLST, Radboud University.

Beyond Aggregate Scores: A Diagnostic Analysis of Portuguese LLM Evaluation Suites

Gustavo Hochgraf[1], **Gabriel Assis**[2], **Aline Paes**[2], **Fabio G. Cozman**[1]

[1]Escola Politécnica, Universidade de São Paulo, São Paulo, SP, Brasil

[2]Instituto de Computação, Universidade Federal Fluminense, Niterói, RJ, Brasil

gusthoch@gmail.com, assisgabriel@id.uff.br, alinepaes@ic.uff.br, fgcozman@usp.br

Abstract. *Benchmarks play a central role in evaluating large language models (LLMs), providing standardized comparisons across models, tasks, and adaptation strategies. However, aggregate benchmark scores often compress performance into a single number, obscuring important variation across tasks. This issue is especially relevant for less-resourced languages like Portuguese, where benchmarks may combine native tasks with translated datasets spanning heterogeneous categories and subareas. In this work, we examine this issue using PoETa v2, a broad Portuguese evaluation benchmark, as a diagnostic setting to evaluate three Qwen3 1.7B variants under different Portuguese adaptation settings. Our results show that, although overall scores differ by less than one percentage point across models, disaggregated analyses reveal substantially different patterns across task origin, category, and subarea. In particular, gains on native Portuguese tasks may coexist with losses on translated tasks, while different adaptation corpora redistribute performance in distinct ways. Taken together, our findings highlight the importance of complementing aggregate scores with multi-level analyses when evaluating Portuguese LLMs, particularly in the context of language-specific adaptation.*

1. Introduction

In recent years, large language models (LLMs) have become a central component of Natural Language Processing (NLP), reshaping how language technologies are developed, adapted, and evaluated [Paes et al. 2026, Real et al. 2026]. As these models are increasingly used across different domains and linguistic contexts, their evaluation has become essential for understanding not only their overall capabilities, but also their limitations and patterns of behavior [Real et al. 2026]. Benchmarks play an important role in this process, since they provide standardized points of comparison across models and adaptation strategies. However, model performance is not uniform across languages, domains, or cultural contexts, and this variation is particularly relevant for languages that remain less represented in the development and evaluation pipelines of widely used LLMs [Huang et al. 2023].

A common practice in LLM evaluation is to compare models through a single aggregate score [Rodriguez et al. 2021]. Although this form of reporting is convenient, it can hide substantial variation across tasks and lead to incomplete or even misleading conclusions about model behavior. Two models with similar overall scores may display very different strengths and weaknesses depending on the task type, the origin of the dataset, or the linguistic and cultural phenomena being evaluated. This is especially important in multilingual and language-specific evaluation, where benchmarks may combine native

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 209–217
October 19-22, 2026 ©2026 Association for Computational Linguistics

tasks, translated datasets, and heterogeneous task categories under the same final score. In such cases, aggregate scores may suggest that two models behave similarly, even when their performance is redistributed in substantially different ways across the benchmark.

Building on this motivation, we examine how aggregate scores may obscure qualitatively different performance patterns in Portuguese LLM evaluation. We instantiate this discussion with PoETa v2 [Almeida et al. 2025b], a benchmark composed of 44 tasks that covers different linguistic and cultural dimensions of Portuguese. This setting allows us to move beyond a single overall number and inspect how results vary according to task origin, category, and subarea. As a case study, we analyze three Qwen3 1.7B [Yang et al. 2025] variants under different Portuguese adaptation settings and investigate whether disaggregating PoETa results changes the interpretation suggested by the aggregate score. This analysis therefore frames aggregate scores as useful but insufficient summaries, especially when evaluating the effects of language-specific adaptation.

2. Related Work

This section provides a brief overview of prior work on Portuguese LLM benchmarks and language-specific adaptation.

2.1. Benchmarks in Portuguese

PoETa is an early benchmark for evaluating LLMs in Portuguese [Pires et al. 2023]. Its second version, PoETa v2 [Almeida et al. 2025b], extends this effort to 44 tasks and more than 20 models, highlighting how computational scale and language-specific adaptation affect performance in Portuguese. Its taxonomy includes 12 tags that may overlap across tasks, and the benchmark also supports analyses by task origin, distinguishing native and translated tasks, as well as by broader capability groupings.

Other Portuguese benchmarks target specific domains or evaluation settings, reinforcing the importance of analyses that go beyond a single aggregate score [Chaves Rodrigues 2023, Severino et al. 2025, Assis et al. 2025]. How aggregate scores are computed also varies across leaderboards: some report a single mean, while others split results by capability or origin, as PoETa v2 does. Some studies have argued that single-number metrics can hide important failures and proposed reporting results in more disaggregated ways [Real et al. 2026, Rodriguez et al. 2021].

2.2. Portuguese-Specific LLM Adaptation

Continued pre-training can improve target-language performance but may also induce forgetting [Gururangan et al. 2020]. In Portuguese, this strategy has been explored through continued pre-training and instruction tuning [Larcher et al. 2023, Almeida et al. 2025a, Garcia et al. 2024, Cruz-Castañeda e Amadeus 2025]. Since these efforts are often assessed with Portuguese-centered benchmarks, how benchmark scores are aggregated becomes a central interpretive issue.

3. Methodology

This section describes the experimental setup used to analyze how different Portuguese adaptation settings affect model performance on PoETa v2. We first present the model variants and corpora considered in the study, and then describe the benchmark subset, evaluation protocol, and reporting scheme used throughout the experiments.

Table 1. Variants evaluated and continued pre-training characteristics.

Variant	Continued pre-training	Corpus	Tokens
Qwen3 1.7B Base	None	–	–
+GigaVerbo adapted	1 epoch	GigaVerbo	∼100 B
+Carolina adapted	1 epoch	Carolina	∼200 M

3.1. Models and Corpora

We evaluate three variants of Qwen3 1.7B, summarized in Table 1. We selected this base model because, within its model-size tier, it combines the strongest previously reported PoETa [Almeida et al. 2025b] results with compatibility with our available computational resources. The first variant is the original Qwen3 1.7B checkpoint. The second is a variant tuned to the GigaVerbo corpus [Corrêa et al. 2025], a large-scale Portuguese collection assembled from multiple publicly available sources and dominated by native Portuguese text. The third is a variant tuned to the Carolina corpus [Crespo et al. 2023], a substantially smaller resource composed of texts in contemporary Brazilian Portuguese with varied typology. In both adaptation settings, training starts from the same Qwen3 1.7B model [Yang et al. 2025] and applies one additional epoch over the respective corpus. We deliberately fix the base model and vary only the adaptation corpus so that observed differences can be attributed to corpus choice rather than to architectural or pre-training variation across distinct baselines; the 1.7B size keeps continued pre-training tractable under our computational budget. This design allows us to compare how the benchmark is interpreted under different adaptation settings.

3.2. Evaluation Setup and Reporting

Following [Almeida et al. 2025b], we first report the benchmark's aggregate metrics, **All**, **Native**, and **Translated**. These correspond, respectively, to the average across all evaluated tasks, the subset of native Portuguese tasks, and the subset of translated tasks. In our experiments, all three aggregates are computed over a fixed public subset of 40 tasks from PoETa v2[1]. We do not impute or reweight missing tasks; therefore, the aggregate values reported in this paper should be interpreted with respect to this 40-task subset rather than to the original 44-task composition of PoETa v2.

The benchmark includes tasks evaluated with accuracy (`acc`), textual similarity (`pearson`), exact match for free-form answers (`best_em`), and F1. Our experiments follow the few-shot setting with dynamic example selection, in which each test instance is evaluated with a small set of automatically selected in-context examples from the same task. To keep evaluation computationally manageable while maintaining coverage across tasks, we limit each task to 300 samples.

Beyond the standard aggregate view, we introduce an additional diagnostic layer by grouping the selected tasks into six mutually exclusive categories: (i) Brazil/exams/culture, (ii) text understanding/QA/classification, (iii) social/safety, (iv) reasoning, (v) math, and (vi) code/other. This grouping is not intended to replace the original PoETa v2

[1]The excluded tasks are `poscomp_greedy`, `arc_challenge_greedy_pt`, `arc_easy_greedy_pt`, and `ethics_commonsense_test_hard_greedy`. These tasks appear in the original full benchmark configuration but are not included in the public subset available.

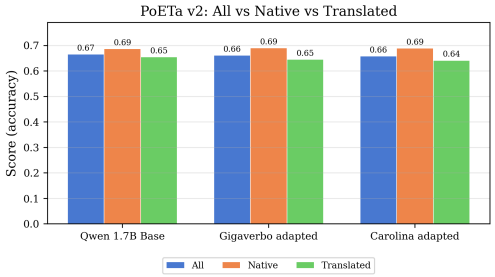


Figure 1. Normalized aggregate scores per variant.

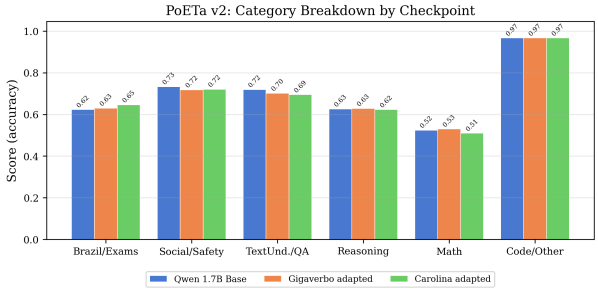


Figure 2. Average score by category for each variant.

taxonomy. Rather, it collapses the benchmark's overlapping tags and, when needed, exam-specific metadata into one primary reporting group per task, allowing us to summarize results at an intermediate level between the overall score and individual tasks. Each task also retains its original native/translated label.

For the Brazil/exams/culture group, we also report results by subarea. Model variants are compared using deltas in percentage points (p.p.) relative to the base model. Together, these category-, subarea-, and delta-based summaries constitute the main analytical extension adopted in this work.

To quantify sampling uncertainty, we accompany reported differences with 95% confidence intervals from a paired nonparametric bootstrap. Because every variant is evaluated on the same 300-example subset of each task, we resample examples with replacement (10,000 replicates) using identical indices for the three variants and recompute task scores and aggregates on each replicate, so deltas are paired at the example level. For the three tasks without recoverable per-example scores (BLUEX, FaQuAD, and MKQA) we use a normal approximation based on the harness-reported standard errors, with unpaired deltas; a difference is significant when its 95% CI excludes zero.[2]

4. Results and Discussion

4.1. Aggregate scores

Figure 1 presents the aggregated scores (All, Native, Translated) for the three variants. We observe that the All score varies by only 0.75 p.p. across models. In contrast, the Native and Translated aggregates move in opposite directions, and the bootstrap separates signal from noise in this movement: the losses on translated tasks are statistically robust (-0.93 p.p., 95% CI $[-1.69, -0.16]$, for +GigaVerbo adapted and -1.32 p.p., $[-2.27, -0.36]$, for +Carolina adapted), whereas the corresponding native gains ($+0.4$ p.p., $[-0.72, +1.47]$, and $+0.29$ p.p., $[-1.00, +1.56]$) are small relative to sampling uncertainty and not individually significant, even though specific native tasks and categories do shift significantly (Sections 4.2 and 4.3). The aggregate score therefore conceals statistically significant degradation on translated tasks behind an apparently stable overall number.

[2]Per-task scores, the task-to-category mapping, and the bootstrap analysis code are available at https://github.com/GustavoHochgraf/brasa-eval.

4.2. Analysis by category

Figure 2 shows the average scores by category. The variant adapted with the Carolina corpus stands out in Brazil/exams/culture (+2.33 p.p., 95% CI $[+0.64, +4.03]$, vs. the base), which is consistent with the stronger presence of Brazilian-oriented content in that corpus. In turn, the GigaVerbo-adapted model shows modest advantages in reasoning and math, a pattern compatible with the scale and heterogeneity of GigaVerbo. In contrast, the base model performs better in text understanding/QA/classification, which aggregates reading comprehension, NLI, sentiment, and fact-verification tasks and in social/safety, which combines hate-speech detection with bias and ethical-reasoning tasks. This indicates that the relative gains from adaptation are not uniform across categories. We note that both of these groupings are intentionally broad to preserve enough tasks per group for stable averages; a finer subdivision would likely surface additional within-category patterns and is left as future work. Under the bootstrap, the Brazil/exams/culture gain of +Carolina adapted and the losses of both adapted variants in text understanding/QA/classification (-1.76 p.p., $[-2.99, -0.51]$, and -2.45 p.p., $[-3.97, -0.98]$) are significant, as is the +GigaVerbo adapted loss in social/safety (-1.48 p.p., $[-2.90, -0.07]$); the movements in reasoning and math are not.

4.3. Largest task-level shifts

Table 2 lists the five largest improvements and drops of each variant relative to the base model. These task-level shifts are particularly revealing because the two adapted variants were exposed to different additional corpora during continued pretraining: GigaVerbo is much larger and more heterogeneous, whereas Carolina is smaller and more centered on contemporary Brazilian Portuguese. For the GigaVerbo-adapted model, the largest gains appear in reasoning tasks (BB Empirical Judgments, BB Mathematical Induction) and in one Brazilian-culture task (BRoverbs History→Proverb), while the main losses are observed in translated reading-comprehension tasks such as BoolQ and IMDb. For the Carolina-adapted model, the largest improvements occur in BB Empirical Judgments and ENEM 2022, which align with the stronger Brazilian orientation of those corpora, whereas the largest losses again occur in translated tasks such as IMDb and StoryCloze. Most of the largest losses fall on translated tasks, with IMDb being the only one to appear in both lists. A possible explanation is that these tasks rely on linguistic cues (e.g., sentiment in IMDb, pronoun resolution in WSC-285) that translation can distort. The results also indicate that corpus alignment may be more influential than corpus size for certain task families, as the smaller but more targeted Carolina corpus yields comparable or stronger gains in Brazilian-specific tasks. Taken together, these task-level results reinforce the paper's main point: different adaptation settings may lead to different performance distributions even when aggregate scores remain close.

4.4. Drill-down on sub-areas ENEM/BLUEX

To better understand the *Brazil/exams/culture* category, we analyzed the ENEM 2022 and BLUEX exams by subarea. Figure 3 shows the average scores for each subarea. In ENEM 2022, the largest gains appear in *Natural Sciences* and *Mathematics*, especially for the Carolina adapted variant (+11.5 p.p. and +9.1 p.p.). In contrast, the *Languages* subarea (which includes Portuguese plus an English or Spanish reading test) declines after Portuguese adaptation (–6.1 p.p. for GigaVerbo adapted), highlighting trade-offs in

Table 2. Top 5 improvements and losses by task relative to Qwen3 1.7B Base (p.p.). † 95% paired-bootstrap CI excludes zero.

GigaVerbo adapted		Carolina adapted	
Task	**Δ**	**Task**	**Δ**
Biggest improvements			
BB Empirical Judgments	+6.1	BB Empirical Judgments	+8.1†
BB Mathematical Induction	+5.7	ENEM 2022	+4.2
BRoverbs History→Proverb	+3.7†	BRoverbs History→Proverb	+3.7
Balanced COPA	+3.3†	BRoverbs Proverb→History	+3.7
BB Analogical Similarity	+2.3	Math MC	+3.3
Biggest losses			
BoolQ	–14.7†	IMDb	–7.3†
AGIEval SAT Math	–3.6	BB Mathematical Induction	–5.7
IMDb	–3.3†	StoryCloze	–5.7†
WSC-285	–3.2	FaQuad	–5.0
BB BBQ	–3.0†	BB BBQ	–4.7†

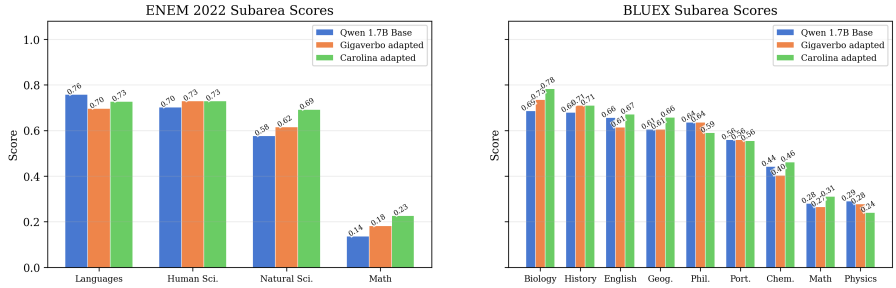


Figure 3. ENEM 2022 and BLUEX scores by subarea, averaged across tasks.

bilingual capability. In BLUEX, we observe consistent gains in *Biology*, *History*, and *Geography*, whereas quantitative subareas such as *Physics* and *Chemistry* show mixed results. Given the small number of questions per subarea, we treat this drill-down as exploratory and do not attach significance claims to subarea-level differences.

Taken together, these results show that aggregate reporting alone is insufficient for interpreting benchmark behavior in Portuguese LLMs. In our case study, models that remain close under **All** display substantially different performance redistributions once the benchmark is broken down by origin, category, subarea, and task. This is precisely why diagnostic reporting is more informative than relying only on a single summary score.

5. Conclusion

This paper presented a diagnostic evaluation of three Qwen3 1.7B variants with PoETa v2 at multiple levels of granularity. We showed that disaggregating results by origin, category, and subarea reveals trade-offs that are not visible under the aggregate score alone. In particular, the adapted variants obtain gains on native tasks while losing performance on translated tasks, with distinct redistribution patterns depending on the adaptation corpus. These findings reinforce the idea that aggregate scores provide a useful starting point for Portuguese LLM evaluation, but should not serve as the main basis for interpretation. Looking ahead, they support treating PoETa v2 less as a single benchmark and more as a set of complementary diagnostics, characterizing each model's capability profile (e.g., robustness to translation or domain sensitivity) beyond its average score.

Acknowledgements

We thank the anonymous reviewers for their constructive feedback, and the BR-LLM group at the Center for Artificial Intelligence (C4AI-USP) for the computational infrastructure and discussions. The C4AI is supported by the São Paulo Research Foundation (FAPESP, grant #2019/07665-4) and by the IBM Corporation. F. G. C. is partially supported by CNPq Pq grant #305753/2022-3. G. A. and A. P. also acknowledge support from the CNPq, grant #307088/2023-5; FAPERJ, processes SEI-260003/002930/2024 and SEI-260003/000614/2023; and the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES), Finance Code 001.

Limitations

This study has some limitations. It is restricted to Qwen3 1.7B variants obtained after one epoch of continued pretraining on two Portuguese corpora, and we do not attempt to isolate the causal effect of each training variable. Sampling uncertainty is quantified with paired bootstrap confidence intervals over 300-example subsets; at this sample size several aggregate differences, including the Native gains, are not individually significant, and the subarea analysis remains exploratory. The analysis also depends on the coverage and design choices of PoETa v2 itself.

Ethics Statement

This research does not involve human subjects. All models were evaluated on public tasks from the PoETa v2 benchmark, and the corpora used in continued pretraining (GigaVerbo and Carolina) are publicly available. The main ethical issue addressed in this work concerns evaluation practice: aggregate scores can obscure meaningful losses in specific task groups and may therefore mislead decisions about model development, comparison, and release. In this sense, more diagnostic reporting can contribute to more transparent assessments of Portuguese LLMs. We also note that both benchmarks and training corpora may reproduce social, cultural, and linguistic biases present in their source data.

Use of Generative Artificial Intelligence

Generative AI tools (Claude, Anthropic) were used to assist with grammatical revision, code generation for data analysis, and LaTeX manuscript formatting. All scientific content, analysis, and interpretation of results are the responsibility of the authors.

References

Almeida, T. S., Nogueira, R., e Pedrini, H. (2025a). Curió-edu 7b: Examining data selection impacts in llm continued pretraining. *arXiv preprint arXiv:2512.12770*.

Almeida, T. S., Pires, R., Abonizio, H., Nogueira, R., e Pedrini, H. (2025b). PoETa v2: Toward more robust evaluation of large language models in Portuguese. *arXiv preprint arXiv:2511.17808*.

Assis, G., Freitas, C., e Paes, A. (2025). Exploring brazil's llm fauna: Investigating the generative performance of large language models in portuguese. *Journal of the Brazilian Computer Society*, 31(1):939–971.

Chaves Rodrigues, R. (2023). Lessons learned from the evaluation of Portuguese language models. Master's thesis, University of Malta.

Corrêa, N. K., Sen, A., Falk, S., e Fatimah, S. (2025). Tucano: Advancing neural text generation for Portuguese. *Patterns*, 6(1):101325.

Crespo, M. C. R. M., Rocha, M. L. d. S. J., Sturzeneker, M. L., Serras, F. R., Mello, G. L. d., Costa, A. S., Palma, M. F., Mesquita, R. M., Finger, M., Sousa, M. C. P. d., Namiuti, C., e Monte, V. M. d. (2023). Carolina: A general corpus of contemporary Brazilian Portuguese with provenance, typology and versioning information.

Cruz-Castañeda, W. A. e Amadeus, M. (2025). Large languages models in brazilian portuguese: A chronological survey. *Journal of the Brazilian Computer Society*, 31(1):1167–1186.

Garcia, G. L., de Medeiros, I. S., e de Oliveira, V. G. (2024). Introducing bode: A fine-tuned large language model for portuguese prompt-based tasks. *arXiv preprint arXiv:2401.02909*.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., e Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

Huang, H., Tang, T., Zhang, D., Zhao, X., Song, T., Xia, Y., e Wei, F. (2023). Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In Bouamor, H., Pino, J., e Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.

Larcher, C., Piau, M., Finardi, P., Gengo, P., Esposito, P., e Caridá, V. (2023). Cabrita: Closing the gap for foreign languages. *arXiv preprint arXiv:2308.11878*.

Paes, A., Vianna, D., e Rodrigues, J. (2026). Modelos de linguagem. In Caseli, H. M. e Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, volume 2, book chapter 1. BPLN, 4 edition.

Pires, R., Abonizio, H., Almeida, T. S., e Nogueira, R. (2023). Sabiá: Portuguese large language models. *arXiv preprint arXiv:2304.07880*.

Real, L., Carvalho, A., e Silva, A. S. d. (2026). Avaliação de grandes modelos de linguagem: Fundamentos, métodos tradicionais e desafios atuais. In Caseli, H. M. e Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, volume 2, book chapter 4. BPLN, 4 edition.

Rodriguez, P., Barrow, J., Hoyle, A., Lalor, J. P., Jia, R., e Boyd-Graber, J. (2021). Evaluation examples are not equally informative: How should that change NLP leaderboards? In Zong, C., Xia, F., Li, W., e Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4486–4503, Online. Association for Computational Linguistics.

Severino, J. V. B. et al. (2025). Benchmarking open-source large language models on Portuguese revalida multiple-choice questions. *BMJ Health Care Informatics*, 32(1):e101195.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., e Qiu, Z. (2025). Qwen3 technical report.

Balancing Coherence and Granularity in Legal Topic Modeling: A BERTopic-based Study on Portuguese Appellate Court Decisions

Marjory S. S. Lima[1], Felipe R. Ahad[1], Vih A. S. Silva[1], Thiago M. Ventura[1], Josiel M. Figueiredo[1], Allan G. de Oliveira[1]

[1]Instituto de Computação, Universidade Federal de Mato Grosso, Brazil

{marjory.lima, felipe.ahad, vitoria.silva7}@sou.ufmt.br,
{thiago, josiel, allan}@ic.ufmt.br

Abstract. *The growing backlog of the Brazilian Judiciary demands scalable semantic organization beyond keyword-based retrieval. This study investigates the trade-off between topic coherence and thematic granularity in neural topic modeling for legal texts. We cluster 1,831 Brazilian labor summaries using BERTopic with LegalBERT-pt embeddings and evaluate the impact of the min_topic_size parameter on coherence (C_v), granularity, and diversity. We propose a balanced evaluation score integrating coherence and granularity. Results show higher coherence ($C_v \approx 0.78$) than Latent Dirichlet Allocation ($C_v \approx 0.49$), while enabling controlled topic resolution. These findings provide a practical criterion for tuning topic models in large-scale legal systems.*

1. Introduction

The Brazilian Judiciary has been facing a significant increase in procedural demand. Over 39,4 million new lawsuits were filed in 2024, raising the national caseload to more than 80 million pending cases [Conselho Nacional de Justiça 2025]. This scenario poses substantial challenges to the management of legal knowledge. The high complexity and volume of judicial documents require increasingly efficient methods for the interpretation, retrieval, and reuse of legal information.

Among the various types of procedural documents, appellate court decisions (*acórdãos*) stand out, as they are collective rulings issued by courts that consolidate legal interpretations on recurring or complex cases. The search for case law, previous decisions that guide the interpretation of legal norms, often depends on the ability to locate thematically similar *acórdãos*, which demands considerable effort from legal professionals in terms of reading, interpreting, and comparing texts. Depending on the context, legal professionals may need to review dozens or even hundreds of *acórdãos* to identify semantic patterns. Although keyword-based search systems are available on court websites, these tools are generally limited to literal term queries and do not account for the semantics or context of the words used. In this context, the need for automated methods to assist in the exploratory analysis of large amounts of legal texts becomes evident, particularly for pattern extraction, information synthesis, and thematic organization of court decisions.

In recent years, there has been significant progress in the application of Natural Language Processing (NLP) techniques in the legal domain, with particular emphasis on topic modeling, contextual embeddings, and semantic clustering mechanisms [da Silva et al. 2024, Silva et al. 2024]. Studies have shown that these techniques can be successfully applied to tasks such as archive organization, case retrieval, and support for drafting legal documents [Amorim et al. 2022, Silveira et al. 2023, Novaes et al. 2023].

While recent research has already explored topic modeling in the legal domain, especially for organizing and retrieving case law, there is still limited emphasis on systematically evaluating the thematic coherence of the generated topics, particularly in unsupervised settings.

By focusing on Brazilian court decisions written in Portuguese, this study investigates how modern NLP techniques can support the semantic organization of legal archives, paying special attention to the impact of modeling choices on interpretability and thematic granularity. However, most evaluation approaches emphasize semantic coherence, neglecting aspects such as diversity and distribution of topics across the corpus. Models optimized exclusively for coherence tend to generate few broad and dominant topics, while increasing granularity can produce fragmented clusters. To address this gap, this work proposes a composite evaluation approach that integrates semantic coherence and topic distribution (as a proxy for granularity), providing a more comprehensive criterion for validating the quality of topics in legal corpora.

2. Related Works

The application of NLP techniques to the legal domain has emerged as a growing research area, driven by the need to organize large volumes of court decisions and assist in the retrieval of similar cases [Martins et al. 2021]. Among the approaches adopted, notable ones include topic modeling techniques, semantic representation using embeddings, document clustering, and textual similarity measurement.

Initially, probabilistic topic models such as Latent Dirichlet Allocation (LDA) [Blei et al. 2003] dominated the literature by enabling the automatic identification of latent themes in large text collections. However, these models have limitations in scenarios involving short documents or specialized vocabulary, as is often the case with court decisions and legal headnotes. To address these limitations, several approaches have been proposed, comparing LDA variants with more recent embedding-based methods. Amorim et al. [2022], for instance, conducted an experimental evaluation using short Twitter texts, showing that models like BERTopic and Pseudo-document based Topic Model (PTM) produce more coherent and interpretable topics than traditional LDA.

Although topic coherence is widely used to evaluate topic models, it does not capture other important aspects of topic quality, such as diversity and document coverage [Röder et al. 2015, Mimno et al. 2009]. Empirical studies show that coherence-based evaluation may favor models with fewer and more dominant topics, limiting thematic granularity [Campagnolo et al. 2022]. In the Brazilian legal context, advances in domain-adapted embeddings have improved semantic modeling, yet evaluation remains largely coherence-driven [Silveira et al. 2023]. This limitation motivates evaluation strategies that jointly consider coherence and diversity. Accordingly, this work proposes a composite score that balances these dimensions to better assess topic quality in legal corpora.

In the Brazilian legal context, the semantic representation of legal documents has also advanced with the development of specialized embeddings. Silveira et al. [2023] introduced LegalBERT-pt, a model based on the BERT architecture, trained on a large national legal corpus, and demonstrated its superiority over generic models in tasks such as sentence classification and entity extraction. This research line is crucial

for the effectiveness of semantic clustering and topic modeling techniques that depend on domain-adapted vector representations.

In parallel, research has emerged focusing on the organization and retrieval of case law through thematic clustering. Borges et al. [2025] explored different clustering algorithms using BERTimbau embeddings applied to Portuguese texts, validating the feasibility of unsupervised methods for thematic segmentation of documents. Complementarily, Novaes et al. [2023] proposed a two-stage architecture for legal case retrieval: first, filtering documents based on topics extracted with BERTopic; then, re-ranking the results using semantic ranking functions. This strategy was evaluated in the Competition on Legal Information Extraction and Entailment (COLIEE 2023) international competition, showing promising results for documents from the Brazilian Superior Court of Justice (STJ).

Another important research line is the measurement of semantic similarity between legal documents, which has gained momentum with the use of large language models (LLMs). Hussain et al. [2023] proposed the CoT-STS method, which decomposes the task of textual comparison into interpretable dimensions (topic, agents, actions, context), enabling more accurate semantic evaluations even in zero-shot scenarios. Although not specific to the legal domain, the method provides an interpretable approach that can be applied to comparisons between *acórdãos* and their summaries.

Overall, the literature points to a convergence around three methodological pillars: (i) the use of specialized semantic representations; (ii) the adoption of unsupervised techniques for thematic clustering; and (iii) the use of human or structural resources, such as case summaries or procedural classes, as validation references. The present work fits within this context by employing legal-specific embeddings and BERTopic for clustering judicial summaries. In addition, it contributes by proposing a composite evaluation approach that integrates semantic coherence and topic distribution, addressing limitations of traditional coherence-based evaluation, which often overlooks the balance and coverage of topics across the corpus.

3. Materials and Methods

3.1. Data Collection and Corpus Description

For the development of this work, 1,831 *acórdãos* made publicly available on the website of the Regional Labor Court of the 23rd Region (TRT-23) were used. The data collection was done through web scraping, a technique widely adopted when data cannot be accessed through an API [Mitchell 2024]. This approach consists of using a tool that systematically and automatically simulates browsing and extracting information from web pages [Glez-Peñaa et al. 2014].

At this stage, to ensure records with standardized structures, the following terms were applied to the text filter fields of the TRT-23 search interface[1]: "*isto posto*" (therefore), for the field "*Contendo as palavras (e)*"; and "*ementa*" (summary), for the field "*Palavras na ementa (e)*". In addition, the document types were limited to court decisions, and the judging bodies were restricted to the "*1ª Turma*", "*2ª Turma*" and

[1] https://pje.trt23.jus.br/jurisprudencia

"*Tribunal Pleno*". These criteria aimed to collect only finalized appellate decisions with clearly delimited summaries.

The decision to rely exclusively on case summaries (*ementas*) was guided not only by computational constraints, but also by their functional role within Brazilian legal practice. *Ementas* are designed to synthesize the core legal issues and outcomes of judicial decisions in a standardized and comparable format, making them particularly suitable for large-scale thematic analysis and jurisprudential mapping. While full decisions (*acórdãos*) provide richer argumentative and doctrinal detail, they often exceed the maximum token limits supported by transformer-based models such as BERT and LegalBERT-pt, requiring truncation or segmentation strategies that may disrupt contextual coherence. By focusing on *ementas*, this study prioritizes consistency, comparability, and practical applicability, while acknowledging that certain nuances of judicial reasoning may not be fully captured.

3.2. Topic Modeling Pipeline

3.2.1. Embedding Generation and Topic Modeling

Semantic representations of the summaries were obtained using pre-trained transformer-based embeddings derived from the BERT (Bidirectional Encoder Representations from Transformers) architecture. For this purpose, LegalBERT-pt [Silveira et al. 2023] was employed, a model specialized in the legal domain and based on BERTimbau [Souza et al. 2020], which stands out as the state of the art among models specialized in Brazilian Portuguese [Silva et al. 2024].

The embeddings were then used as input for the topic modeling phase, an approach that enables the synthesis and organization of large document corpora, allowing the discovery of themes and their relationships. This process began with a dimensionality reduction step using UMAP [McInnes et al. 2018], which projects high-dimensional embeddings into a dense representation optimized for neighborhood preservation. The resulting reduced embeddings were then modeled using BERTopic [Grootendorst 2022], a method that employs the HDBSCAN [Campello et al. 2013] algorithm to identify clusters, which are subsequently represented as topics by comparing the importance of words using a class-based variant of the TF-IDF measure, called c-TF-IDF (1).

$$Wt,\, c = TF_{t,\, c} \cdot log\left(1 + \frac{A}{F_t}\right) \tag{1}$$

As shown in Equation 1, $TF_{t,\, c}$ refers to the frequency of term t in class c, F_t refers to the frequency of term t across all classes, and A represents the average number of words per class.

3.2.2. Experimental Configurations with BERTopic

To systematically evaluate the behavior of neural topic modeling under different representational and granularity settings, multiple experimental configurations were executed using the BERTopic framework. In all configurations, the same document embeddings, UMAP dimensionality reduction model, tokenized documents, and dictionary were employed, ensuring that performance differences could be attributed solely to modeling choices rather than data variation.

The default configuration served as a reference point and employed unigram-based representations only, with a fixed minimum topic size of 10 documents. This setup reflects the standard usage of BERTopic without explicit modeling of multiword expressions. To assess the impact of compound legal expressions on topic quality, two additional configurations were evaluated using expanded n-gram representations. The bigram configuration incorporated n-grams in the range (1,2), while the trigram configuration extended this range to (1,3). In both cases, the minimum topic size was kept constant at 10 documents to isolate the effect of n-gram expansion on topic coherence and granularity.

Beyond fixed configurations, a parametric exploration of topic granularity was conducted by systematically varying the *min_topic_size* parameter from 2 to 52. This interval was empirically defined based on preliminary experiments, which indicated that values outside this range led to either excessively fragmented topics or overly dominant clusters. Parameter selection followed empirical tuning guided by prior studies using BERTopic in short-text corpora [Amorim et al. 2022; Novaes et al. 2023]. For each value in this range, a separate BERTopic model was trained and evaluated, enabling an empirical analysis of how topic size constraints influence coherence, diversity, and document distribution across topics.

From this parametric sweep, two reference configurations were identified. The coherence-optimal configuration corresponds to the model achieving the highest C_v, while the balanced configuration was selected based on the maximum value of the balanced score, which jointly considers semantic coherence and topic distribution equity. These configurations represent, respectively, the upper bound of semantic consistency and the best trade-off between coherence and thematic granularity.

3.3. Evaluation Approach

3.3.1. Quantitative Metrics

Finally, to evaluate the quality of the generated topics, the Coherence Score (C_v) [Mimno et al. 2011] was applied. This is a metric that uses a variation of the Normalized Pointwise Mutual Information (NPMI) [Bouma 2009] measure, in which the co-occurrence of a word within a given topic is calculated by comparing that word with all the other words in the same topic. This metric stands out for its high sensitivity to the introduction of noise in topics, making it particularly effective in identifying semantically consistent groupings [Campagnolo et al. 2022]. This is an essential feature in the legal domain, where topics must reflect precise and well-defined normative concepts, such as "*horas extras*" (overtime) or "*assédio moral*" (moral harassment) even in an unsupervised setting. If we consider the words "*horas*" (hours) and "*extras*" the C_v calculation begins with the creation of a vector for both words, given the Equation 2:

$$\nu_{horas,\, extras} = NPMI(horas,\, extras)^{\gamma} = \left(\frac{PMI(horas, extras)}{-\log(P(horas, extras) + \epsilon)}\right)^{\gamma} \quad (2)$$

Then, for each word in the topic, a vector is created containing its NPMI with all other words in the same group. For example, the vector for "*horas*" includes its co-occurrence with "*extras*" and "*trabalhadas*" (3).

$$\vec{v}_{horas} = \{\, NPMI(horas,\, horas)^{\gamma} \,,\, NPMI(horas,\, extras)^{\gamma} \,,\, NPMI(horas,\, trabalhadas)^{\gamma} \,\} \quad (3)$$

Finally, the distance between the vectors are calculated using the cosine similarity measure (4).

$$C_v = \frac{1}{3} \cdot (cos(\vec{v}_{horas}, \vec{v}_c), cos(\vec{v}_{extras}, \vec{v}_c), cos(\vec{v}_{trabalhadas}, \vec{v}_c)) \quad (4)$$

In addition to the C_v metric, which evaluates the semantic consistency of the most representative terms of each topic, a novel metric is proposed to also consider the balance of document distribution across the topics. The diversity of the topics was defined by Equation 5.

$$Diversity = 1 - \left(\frac{Qty.\ documents\ on\ biggest\ topic}{Total\ no\ of\ documents}\right) \quad (5)$$

Subsequently, the balanced score was defined as:

$$Balanced\ Score = C_v \cdot Diversity \quad (6)$$

While metrics like Topic Diversity [Dieng et al. 2020], the Gini coefficient, and entropy-based measures [Röder et al. 2015] evaluate structural properties independently, the proposed Balanced Score provides a simple, interpretable selection criterion. Rather than replacing these measures, it explicitly couples semantic coherence with topic dominance to penalize overly concentrated clusters without requiring additional parameterization, simultaneously assessing semantic quality and distributional equity to identify the configuration considered the most balanced.

3.3.2. Baseline Model: Latent Dirichlet Allocation

To contextualize the results obtained with neural topic modeling, LDA was adopted as a classical probabilistic baseline [Blei et al. 2003, Mimno et al. 2009, Röder et al. 2015]. The model was trained on the same corpus of tokenized summaries using the same dictionary employed in the BERTopic experiments. Preprocessing additionally included the removal of standard Portuguese stopwords to reduce the influence of non-informative tokens on word co-occurrence statistics.

The number of topics was set to 82, matching the balanced BERTopic configuration to enable a fair comparison. Training was performed with a fixed random seed and 10 passes to ensure convergence stability. Topic quality was evaluated using the same C_v coherence metric adopted for BERTopic. The resulting coherence score (0.4883) serves as a reference for comparing probabilistic and neural topic modeling approaches on Portuguese legal texts.

3.3.3. Qualitative Evaluation by Domain Experts

In addition to the quantitative evaluation metrics, a qualitative assessment was conducted to examine the interpretability and practical relevance of the topics generated by the proposed approach. This evaluation aimed to complement automatic measures by incorporating expert judgment, which is particularly important in the legal domain, where semantic adequacy and doctrinal coherence are essential.

The qualitative analysis focused on the balanced BERTopic configuration, selected for presenting the best trade-off between semantic coherence and equitable document distribution, as indicated by the proposed balanced score. From the total set of 113 topics produced by this configuration, a subset of five topics was selected for expert evaluation. The selected topics were chosen to reflect thematic diversity and to

cover different types of legal issues commonly addressed in labor court decisions. For each topic, evaluators were provided with: (i) a list of the most representative terms automatically extracted by the model, and (ii) a sample of *ementas* associated with the topic, allowing for an informed assessment of the relationship between the topic descriptors and the underlying judicial content.

The evaluation was conducted through an online survey, in which three legal domain specialists were asked to assess each topic based on a single evaluative statement encompassing three dimensions simultaneously: legal thematic coherence, clarity and adequacy of the representative terms, and practical usefulness for legal research or jurisprudential analysis. Responses were recorded using a five-point Likert Scale ranging from "Strongly disagree" to "Strongly agree" [Joshi et al. 2015]. Each evaluator assessed all selected topics independently. The results of this qualitative evaluation were subsequently analyzed to identify patterns of agreement and to contrast expert perceptions with the quantitative findings derived from coherence and distribution-based metrics.

3.4. Reproducibility Considerations

All experimental runs, including BERTopic configurations and the LDA baseline, were executed with fixed random seeds to ensure deterministic behavior. Intermediate results, evaluation metrics, and configuration parameters were systematically logged and stored. The complete experimental pipeline, including preprocessing routines, modeling configurations, and evaluation scripts, is publicly available at https://github.com/msoehn3/legal-topic-modeling, enabling full replication of the results.

4. Results

The experiments were conducted with the objective of evaluating the performance of the BERTopic algorithm to identify latent topics in *ementas* under different parameter configurations and compare them with a LDA baseline. The consolidated results are presented in Table 1, which shows, for each configuration, the number of topics generated, the average thematic coherence (measured by the C_v metric), the proportions of topics with the lowest and highest document concentration, as well as the deviation between these extremes, used as a proxy for thematic balancing.

Table 1 - Results of the Experiments

Experiment	Min. Topic Size	Topics	C_v	Balanced Score	Min. Proportion	Max. Proportion	Deviation
Balanced	4	113	0.75	**0.63**	0.22%	16.82%	**16.60%**
Optimal	7	71	**0.78**	0.62	0.44%	20.32%	19.88%
Default	10	52	0.75	0.57	0.55%	23.48%	22.94%
Trigrams	10	52	0.74	0.57	0.55%	23.48%	22.94%
Bigrams	10	52	0.74	0.57	0.55%	23.48%	22.94%
LDA Baseline	-	113	0.49	-	-	-	-

Among the BERTopic configurations, coherence values ranged from 0.7409 to 0.7782, substantially outperforming the LDA baseline, which achieved a C_v score of 0.4883 with the same number of topics (113). This confirms the advantage of neural

topic modeling with contextualized legal embeddings over probabilistic bag-of-words approaches in the legal domain.

The configuration Optimal exclusively for coherence achieved the highest coherence score (C_v = 0.78), generating 71 topics. While this setting slightly increased semantic consistency, it also exhibited a higher concentration of documents in its largest topic (max proportion = 20.32%), resulting in a larger deviation between the smallest and largest topics (19.88%). This behavior suggests a moderate tendency toward topic concentration when coherence is prioritized, with potential implications for thematic granularity.

The Balanced configuration produced a finer-grained topic structure with 113 topics, while preserving a high level of coherence (C_v = 0.75). This configuration achieved the highest balanced score (0.63) and reduced document concentration across topics, yielding the smallest deviation (16.60%) among the BERTopic variants. As a result indicating a slightly more even distribution of documents with only a marginal reduction in semantic coherence.

The Default configuration generated 52 topics with a coherence of 0.75 and moderate balanced score (0.57). Expanding the representation to include bigrams and trigrams did not improve performance. Both Bigrams and Trigrams configurations resulted in slightly lower coherence values (0.7409 and 0.7443, respectively), identical topic counts, and unchanged distributional characteristics. These results suggest that, for this corpus, increasing the n-gram range did not significantly enhance topic quality.

Figure 1 illustrates the trade-off between coherence and diversity as a function of the *min_topic_size* parameter. The curve shows that coherence increases as topics become larger and more semantically stable, reaching a peak at intermediate values. Beyond this point, diversity declines sharply, reflecting the concentration of documents into fewer dominant topics. This behavior highlights the structural tension between semantic coherence and thematic granularity in legal topic modeling and supports the use of a combined evaluation criterion.

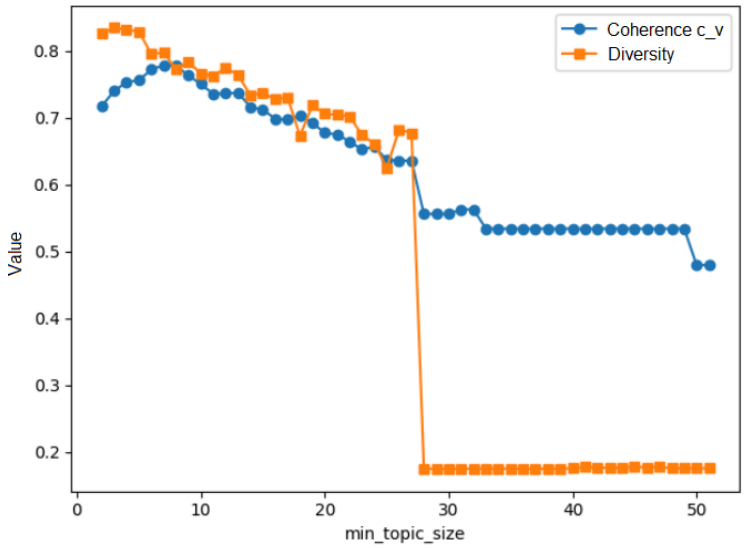


Figure 1. ***Sensitivity to the min_topic_size parameter.***

The qualitative inspection of the LDA topics further reinforces the quantitative findings. As illustrated by the extracted topic-word lists, many LDA topics are

dominated by function words and high-frequency legal terms (e.g., “de”, “da”, “o”, “que”), with limited semantic specificity. Even when legally relevant terms appear (e.g., “horas extras”, “insalubridade”, “dano moral”), they are often mixed with syntactic fillers, reducing interpretability. This lack of semantic focus is consistent with the low coherence score obtained by LDA (0.49) and contrasts sharply with the more thematically cohesive topics produced by BERTopic.

With C_v coherence of up to 0.74, using BERTopic combined with LegalBERT-pt, this study presents robust performance in generating semantically consistent topics, overcoming challenges frequently observed in the literature. When compared to the results reported by Amorim et al. [2022], who obtained C_v coherence values in the range of 0.492 to 0.623 in short texts using different topic modeling techniques, the results of this work are significantly superior. The difference can be attributed mainly to the use of specific embeddings, which capture more robust semantic relationships.

The results of the qualitative evaluation are presented in Table 2. The evaluation conducted with legal experts corroborates the quantitative findings obtained through coherence and distribution-based metrics. Topics associated with well-established legal concepts, such as working hours, discriminatory dismissal, and procedural appeals, achieved the highest levels of expert agreement, indicating strong semantic coherence and practical relevance. Lower variance across evaluators further suggests consistency in expert perception, reinforcing the interpretability and robustness of the proposed topic modeling approach.

Table 2 - Human Evaluation of BERTopic Topics (Balanced Configuration)

ID	Representative Terms	Mean Score	Std. Dev.	Agreement Level
81	*dispensa, discriminatória, discriminação, hiv, portador, preconceito, estigma, grave, justa, discriminatório*	4.33	0.58	High
14	*agravo, petição, decisão, impugnação, conhecido, cálculos, oferecida, ultrapassa, barreira, aprecia*	4.67	0.58	Very High
3	*jornada, ponto, cartões, ônus, controles, extras, horários, 338, horas, prova*	5.00	0.00	Full
94	*71, intervalo, 4º, térmico, suprimido, indenizatória, 384, 2017, redação, 11*	3.67	0.58	Moderate
35	*audiência, comparecimento, arquivamento, acordo, extrajudicial, importa, 844, homologação, partes, inaugural*	4.33	0.58	High

Overall, the results indicate that it is possible to obtain semantically consistent and distributively balanced models. The C_v metric proved useful as an initial evaluation criterion, but its combination with distribution measures provides a more robust analysis for real-world applications in the legal context.

5. Discussion

The results obtained indicate the potential of the proposed approach for identifying thematic structures within judicial documents. The experimental results evidenced a substantial advantage of BERTopic over the LDA baseline in the context of Brazilian labor court decisions. While LDA achieved a low coherence score, BERTopic consistently produced more semantically coherent and interpretable topics across all

evaluated configurations. This gap highlights the limitations of probabilistic topic models based on bag-of-words assumptions when applied to complex legal texts, which are characterized by long documents, formulaic language, and high lexical redundancy. This finding is consistent with prior studies that report reduced effectiveness of LDA in short or domain-specific texts, where limited word co-occurrence hinders topic formation [Amorim et al. 2022].

In addition, the dataset is composed of labor court decisions, which typically involve recurrent thematic patterns across documents. This characteristic provides important context for interpreting the results, especially given the consistent performance differences observed between the evaluated models.

An important finding of this study concerns the trade-off between semantic coherence and topic diversity controlled by the *min_topic_size* parameter. Configurations optimized exclusively for coherence tended to generate fewer, more dominant topics, increasing document concentration and reducing thematic granularity. Conversely, configurations favoring diversity produced a larger number of more evenly distributed topics, at the cost of a modest reduction in coherence.

Furthermore, prior work has emphasized that coherence metrics alone may not fully capture topic quality. Campagnolo et al. 2022 demonstrate that coherence measures are sensitive to noise but do not account for distributional aspects of topics, while Röder et al. 2015 highlight that different coherence formulations may favor different topic structures. In line with these findings, the results of this study provide empirical evidence of this effect. The coherence-optimal configuration, despite achieving the highest C_v score (0.78), also exhibits a higher maximum topic proportion (20.32%) compared to the balanced configuration (16.82%), indicating stronger dominance of specific topics.

Notably, the difference between the Optimal and Balanced configurations was quantitatively small, suggesting a plateau effect in which further gains in coherence no longer translate into meaningful improvements in topic quality. This observation reinforces the limitation of relying solely on coherence metrics, as also discussed in the literature, where coherence improvements may not correspond to better semantic coverage or usability in downstream tasks. Importantly, a high C_v score indicates lexical coherence within the corpus, but does not guarantee identical legal outcomes across documents, reinforcing the necessity of complementary expert analysis.

The alignment between human evaluations and automatic coherence metrics provides empirical evidence that the proposed balanced score effectively captures dimensions of topic quality that are meaningful to legal experts. While configurations optimized solely for coherence may inflate semantic consistency at the expense of thematic granularity, the balanced configuration indicates that it is possible to achieve high interpretability without sacrificing topic diversity. This alignment strengthens the argument that the proposed evaluation framework is suitable for real-world legal information retrieval and jurisprudential analysis tasks.

6. Limitations

This study presents some limitations that should be considered when interpreting the results. First, while standard Portuguese stopwords and digits were removed to prevent non-informative function words from dominating LDA topics, the model was evaluated

using default tokenization parameters without task-specific optimization. Since probabilistic topic models rely on word co-occurrence patterns, this setup may not reflect the best achievable performance of LDA, particularly in short or highly standardized texts. Additionally, while the human evaluation provided key expert validation, it was restricted to a representative subset of five topics due to manual annotation costs, serving as a focused qualitative sample rather than an exhaustive assessment of all generated clusters. Finally, the experiments were conducted on a specific corpus of Brazilian labor court decisions, which may limit the generalizability of the findings to other legal domains or languages, especially those with different linguistic or structural characteristics, which may influence topic distribution patterns.

7. Conclusion and Future Work

This study addressed the challenge of organizing large volumes of judicial documents by investigating how topic modeling can generate semantically coherent and well-balanced thematic structures in Brazilian labor court summaries. The objective was to analyze the trade-off between topic coherence and granularity, and to propose an evaluation framework capable of capturing both aspects.

The results show that BERTopic consistently outperforms the LDA baseline in terms of semantic coherence. However, configurations optimized exclusively for coherence tend to produce dominant topics and reduced thematic coverage. In contrast, the balanced configuration achieved a more equitable distribution of documents while maintaining high coherence, resulting in the best overall performance. These findings indicate that coherence alone is insufficient to assess topic quality and that incorporating distributional aspects leads to more informative and usable topic structures.

For future work, it is recommended to expand the scope of the analysis beyond abstract summaries by incorporating full judicial decisions, which may enable the identification of more diverse and detailed topics. It is also important to validate the proposed approach on datasets from other courts, in order to assess its generalizability across different legal contexts. Additionally, comparing the performance of different domain-specific embedding models may help clarify their impact on topic quality. Finally, exploring Neural Topic Models (NTMs) constitutes a promising direction to further extend the results presented in this study.

Ethics Statement

This research uses publicly available legal documents and does not involve personal or sensitive data. In addition, a qualitative evaluation was conducted with human participants. The data collection was performed anonymously, without the collection of personal or sensitive information. All procedures followed applicable ethical standards.

Use of Generative AI

Generative AI tools were used to assist in language revision and text refinement. All scientific content, experimental design, and analysis were conducted by the authors.

References

Amorim, A., Murrugarra-Llerena, N., Silva, V., de Oliveira, D. and Paes, A. (2022) "Modelagem de Tópicos em Textos Curtos: uma Avaliação Experimental", In: Proceedings of the 37th Simpósio Brasileiro de Banco de Dados (SBBD), Búzios, Brazil. Sociedade Brasileira de Computação. DOI: https://doi.org/10.5753/sbbd.2022.224314.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003) "Latent Dirichlet Allocation", Journal of Machine Learning Research, vol. 3, pp. 993–1022. Available at: https://dl.acm.org/doi/10.5555/944919.944937.

Borges, B. R. (2025) "Comparison of Clustering Techniques in Text Documents in Portuguese", ISys - Brazilian Journal of Information Systems, vol. 18, no. 1, pp. 4:1–4:17. DOI: https://doi.org/10.5753/isys.2025.5029.

Bouma, G. (2009) "Normalized (Pointwise) Mutual Information in Collocation Extraction". Available at: https://svn.spraakdata.gu.se/repos/gerlof/pub/www/Docs/npmi-pfd.pdf.

Campagnolo, J. M., Duarte, D. and Dal Bianco, G. (2022) "Topic coherence metrics: How sensitive are they?", Journal of Information and Data Management, vol. 13, no. 4, pp. 476–491. DOI: https://doi.org/10.5753/jidm.2022.2181.

Campello, R., Moulavi, D. and Sander, J. (2013) "Density-Based Clustering Based on Hierarchical Density Estimates", In: Advances in Knowledge Discovery and Data Mining, Springer. DOI: https://doi.org/10.1007/978-3-642-37456-2_14.

Conselho Nacional de Justiça (CNJ) (2025) Justiça em Números 2025: Ano-base 2024. Departamento de Pesquisas Judiciárias, Brasília, DF.

da Silva, L., Rodrigues, M., Archanjo, A., Pessoa, L., Silva, M., de Almeida, T., & Silveira, L. (2024). Segmentação Textual Baseada em Tópicos em Português Utilizando BERTimbau. In Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana, (pp. 32-36). Porto Alegre: SBC. doi:10.5753/stil.2024.245080.

Dieng, A. B., Ruiz, F. J. R. and Blei, D. M. (2020) "Topic Modeling in Embedding Spaces", Transactions of the Association for Computational Linguistics, vol. 8, pp. 439–453. DOI: https://doi.org/10.1162/tacl_a_00325.

Dragnut, E. et al. (2009) "Stop word and related problems in web interface integration", Proceedings of the VLDB Endowment. DOI: http://dx.doi.org/10.14778/1687627.1687667.

Glez-Peña, D. et al. (2014) "Web scraping technologies in an API world". Briefings in Bioinformatics. DOI: https://doi.org/10.1093/bib/bbt026.

Grootendorst, M. (2022) "BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure", arXiv preprint. DOI: https://doi.org/10.48550/arXiv.2203.05794.

Hussain, M., Rehman, U. U., Nguyen, T. D. T. and Lee, S. (2023) "CoT-STS: A Zero-shot Chain-of-thought Prompting for Semantic Textual Similarity", In:

Proceedings of the 6th Artificial Intelligence and Cloud Computing Conference (AICCC), Kyoto, Japan. ACM. DOI: https://doi.org/10.1145/3639592.3639611.

Joshi, A., Kale, S., Chandel, S., & Pal, D. K. (2015). Likert Scale: Explored and Explained. Current Journal of Applied Science and Technology, 7(4), 396–403. https://doi.org/10.9734/BJAST/2015/14975.

Martins, V. S. and Silva, C. D. (2021) "Text Classification in Law Area: A Systematic Review", In: Proceedings of the Symposium on Knowledge Discovery, Mining and Learning (KDMiLe). DOI: http://dx.doi.org/10.5753/kdmile.2021.17458.

Mimno, D. et al. (2009) "Optimizing Semantic Coherence in Topic Models", In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). DOI: https://dl.acm.org/doi/10.5555/2145432.2145462.

Mitchell, R. (2024) Web Scraping com Python (3rd ed.). São Paulo: Novatec Editora.

Novaes, L. P., Vianna, D. and da Silva, A. S. (2023) "Modelagem de Tópicos para a Tarefa de Recuperação de Casos Legais", In: Proceedings of the 38th Simpósio Brasileiro de Banco de Dados (SBBD), Belo Horizonte, Brazil. Sociedade Brasileira de Computação. DOI: https://doi.org/10.5753/sbbd.2023.232576.

Röder, M., Both, A. and Hinneburg, A. (2015) "Exploring the Space of Topic Coherence Measures", In: Proceedings of the 8th ACM International Conference on Web Search and Data Mining (WSDM), Shanghai, China. ACM, pp. 399–408. Available at: http://svn.aksw.org/papers/2015/WSDM_Topic_Evaluation/public.pdf.

Sarica, S. and Luo, J. (2021) "Stopwords in Technical Language Processing", PLOS ONE, vol. 16, no. 7. DOI: https://doi.org/10.1371/journal.pone.0254937.

Silva, M. et al. (2024) "Evaluating Domain-adapted Language Models for Governmental Text Classification Tasks in Portuguese", In: Proceedings of the 39th Brazilian Symposium on Databases (SBBD), Florianópolis, Brazil. DOI: https://doi.org/10.5753/sbbd.2024.240508.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In Proceedings of the Brazilian Conference on Intelligent Systems, pages 403–417. Springer. DOI: https://doi.org/10.1007/978-3-030-61377-8_28.

Silveira, R., Ponte, C., Almeida, V., Pinheiro, V. and Furtado, V. (2023) "LegalBERT-pt: A Pretrained Language Model for the Brazilian Portuguese Legal Domain", In: Proceedings of the 12th Brazilian Conference on Intelligent Systems (BRACIS), Belo Horizonte, Brazil. Sociedade Brasileira de Computação. Available at: https://sol.sbc.org.br/index.php/bracis/article/view/28420.

Bootstrapping Text Anomaly Detection with LLM-Generated Weak Supervision

Fabio Masaracchia Maia[1], **Anna Helena Reali Costa**[1]

[1]Escola Politécnica, Universidade de São Paulo, São Paulo, Brazil

fabio.masaracchia@usp.br, anna.reali@usp.br

Abstract. *Text anomaly detection is challenging because anomalous instances often share vocabulary and surface form with normal data, making them hard to distinguish without semantic understanding. Semi-supervised methods can significantly outperform unsupervised baselines, but rely on labeled anomalies that are rarely available in practice. LLMs encode rich semantic knowledge that can approximate human judgments, yet using them directly as detectors is costly at inference time and sensitive to prompt design, while generating synthetic outliers risks distribution mismatch with real anomalies. We propose a different strategy: treating a compact, locally deployed LLM as a noisy annotator over real data. The LLM scores a small subset of unlabeled documents once at training time, producing weak labels that refine a sentence encoder via contrastive fine-tuning and train a lightweight downstream detector — without cloud APIs, human annotation, or curated anomaly datasets. Across four datasets spanning two languages and two tasks, the approach recovers up to 79% of the gap to oracle upper bounds while using 14–91× fewer labeled anomalies. We further identify three failure modes that explain when LLM-generated weak supervision succeeds or breaks down.*

1. Introduction

Semi-supervised anomaly detection can dramatically outperform unsupervised methods — but only when labeled anomalies are available [Pang et al. 2021, Xu et al. 2023b]. In text domains, this creates a practical bottleneck: the labels that would most improve detection are precisely the hardest to obtain, because anomalies are rare, diverse, and may share vocabulary and register with normal instances (e.g. hate speech vs. neutral posts on the same platform).

Large Language Models (LLMs) offer a potential solution, as their broad pretraining enables them to approximate human judgments on many NLP annotation tasks [Wang et al. 2021, He et al. 2024]. Prior work in anomaly and out-of-distribution detection has largely explored this potential in two ways. *LLM-as-detector* approaches query the model directly for anomaly judgments, but their predictions can be sensitive to prompt design and often lack calibration across domains [Liu et al. 2024, Yang et al. 2025]. *LLM-as-generator* approaches produce synthetic outlier examples for training, but risk distribution mismatch between generated and real anomalies [Hendrycks et al. 2019, Xu et al. 2023a]. However, neither approach directly yields a lightweight detector that is both stable across domains and inexpensive at inference time.

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 231–244
October 19-22, 2026 ©2026 Association for Computational Linguistics

We take a different approach: instead of using the LLM for detection or data generation, we use it as a *noisy annotator* over real data. A compact, locally deployed model (Qwen2.5-Instruct [Qwen Team 2025], 7B/14B, 4-bit quantization) scores a small subset of *real* unlabeled documents once at training time, producing weak anomaly labels. These labels are used to refine the sentence encoder via contrastive fine-tuning and to train a lightweight downstream detector — with no cloud APIs, no human annotators, and no annotation budget. By annotating real data rather than generating synthetic examples, and by decoupling annotation from inference, the approach avoids both the instability of LLM-based detection and the distribution mismatch of synthetic generation.

Our contributions:

- An LLM-in-the-loop framework that transfers weak semantic supervision into a lightweight text anomaly detector, eliminating the need for human annotation.
- Empirical evidence that the approach closes up to 79% of the gap to oracle upper bounds, using 14–91$\times$ fewer anomaly labels, across two tasks and two languages.
- Identification of three failure modes and the finding that the pipeline tolerates high volumes of false positives but fails under systematic label inversion — a distinction between quantitative and directional annotation noise with practical implications for deployment.

2. Related Work

2.1. Text Anomaly Detection

Anomaly detection is well established for tabular and image data [Pang et al. 2021], but the text domain introduces distinct challenges: documents are high-dimensional, semantically rich, and often require domain-specific knowledge to separate anomalous from normal content [Chandola et al. 2009, Boutalbi et al. 2023]. Classical one-class methods — OC-SVM [Schölkopf et al. 2001], Isolation Forest [Liu et al. 2008] — applied to sparse representations struggle with the semantic similarity between classes. Neural methods learn denser representations: CVDD [Ruff et al. 2019] adapts one-class deep learning to text, and DATE [Manolache et al. 2021] uses self-supervised corruption for fully unsupervised training.

A key finding from Xu et al. [Xu et al. 2023b], who benchmark 22 algorithms across 17 corpora, is that *encoder quality matters more than the choice of AD algorithm*. This motivates our design: rather than proposing a new detection architecture, we focus on improving the encoder via contrastive fine-tuning before detection. Das et al. [Das et al. 2023] confirm that even a handful of labeled anomalies yields large gains, and Ait-Saada & Nadif [Ait-Saada and Nadif 2023] address multi-topic short-text AD, a setting close to ours. Both assume clean human labels are available. A recent comparative study [Maia and Costa 2025] benchmarks semi-supervised text AD across multiple languages, confirming that the label scarcity bottleneck persists across linguistic settings. Our work follows this semi-supervised paradigm but replaces the human annotation assumption with LLM-generated weak labels.

2.2. LLMs for Anomaly and OOD Detection

Recent work explores LLMs as anomaly or out-of-distribution (OOD) detectors. Although the two tasks share the goal of flagging deviations from normality, OOD detection

typically assumes that anomalies form a separable distribution, whereas AD must handle anomalies that overlap semantically with normal data [Chandola et al. 2009] — as in our hate speech setting. Xu & Ding [Xu and Ding 2025] categorise approaches into two main roles: LLMs as detectors (prompting or embedding-based) and LLMs as generators (producing synthetic data or auxiliary labels).

In detection-based approaches, LLMs are queried directly for anomaly judgments. Liu et al. [Liu et al. 2024] and Yang et al. [Yang et al. 2025] find competitive but inconsistent performance: LLM effectiveness is highly task-dependent, with failure modes on fine-grained distinctions and sensitivity to prompt design. This limits their reliability as standalone detectors, especially where stable, calibrated decision boundaries are needed. In generation-based approaches, LLMs enrich training data rather than serving as detectors. Outlier exposure methods introduce proxy anomalies [Hendrycks et al. 2019], and CoNAL [Xu et al. 2023a] generates novel-class examples via prompting for contrastive OOD training. However, these strategies rely on *synthetic* samples whose distribution may not reflect real anomalies, risking distribution mismatch.

2.3. Weak Supervision and Few-Shot Fine-Tuning

Weak supervision refers to learning from imperfect supervisory signals, such as heuristic rules, noisy annotations, or weakly informative labels, instead of fully reliable human-curated ground truth [Ratner et al. 2020, Zhou 2018].Recent work has explored LLM annotation as an alternative to crowdsourcing [Wang et al. 2021, He et al. 2024]. However, these approaches assume *well-defined label spaces* with explicit class boundaries (e.g. sentiment polarity, named topics). Anomaly detection requires the annotator to infer an *implicit normality boundary* — what counts as "normal" for a specific corpus — without labeled examples and without a closed set of target classes. This makes LLM annotation both harder (the prompt must convey a distributional concept) and noisier (the LLM has no ground-truth anchors to calibrate against).

Since encoder quality is the dominant factor in text AD performance [Xu et al. 2023b], refining representations with the few available labels is a natural strategy. Contrastive fine-tuning reshapes the embedding space by pulling instances with the same label closer together while pushing instances with different labels farther apart [Khosla et al. 2020], effectively sharpening the normal–anomaly boundary before any detector is trained. Among few-shot approaches, SetFit [Tunstall et al. 2022] provides a practical implementation of this idea by fine-tuning a sentence transformer from few labeled pairs, achieving strong performance with as few as 8 examples per class.

3. Methodology

3.1. Problem Formulation and Pipeline

We frame the task as *semi-supervised anomaly detection*: a small set of labeled anomalies is available at training time, but the bulk of the data is unlabeled. An important subtlety is that "anomaly" can be defined both as statistically rare (*low-frequency*) and as content deviating from expectations (*semantic*). We construct the low-frequency anomaly by subsampling the minority class to $\sim 5\%$, but the LLM annotator operates semantically, judging content against natural-language criteria.

Let $\mathcal{D}$ be a dataset with a designated normal class. We construct X_{train} with contamination rate $\rho \approx 5\%$ and hold out X_{test} preserving the same proportions. No human-labeled anomalies are available.

Figure 1 illustrates the pipeline. All documents are encoded into dense vectors using the pre-trained multilingual sentence encoder *distiluse-base-multilingual-cased-v2* [Reimers and Gurevych 2019]. A candidate set $S \subset X_{\text{train}}$, $|S| = N \ll |X_{\text{train}}|$, is selected via one of two sampling strategies (Section 3.2). For each candidate $x_i \in S$, a structured prompt is constructed as:

$$P_i = T(x_i,\, d_{\text{task}},\, c_{\text{normal}},\, c_{\text{anomaly}}) \tag{1}$$

where $T(\cdot)$ is the prompt-building function, d_{task} is a task description, and $c_{\text{normal}}, c_{\text{anomaly}}$ define the normality boundary (all dataset-specific). The prompt is structured to provide a clear task definition, explicit criteria for both classes, and a request for a continuous score paired with a one-sentence explanation, reducing ambiguity and promoting consistent label generation across documents. The LLM returns a continuous anomaly score and a one-sentence explanation:

$$(s_i,\, r_i) = f_{\text{LLM}}(P_i), \quad s_i \in [0, 1] \tag{2}$$

The score is binarized at threshold $\tau = 0.5$ to obtain a weak label $\hat{y}_i = \mathbf{1}[s_i \geq \tau]$, yielding the weak anomaly set $\hat{A} = \{x_i \in S : \hat{y}_i = 1\}$. This threshold corresponds to the standard midpoint of the $[0, 1]$ score range; no tuning was performed, as our focus is on the weak supervision strategy rather than threshold calibration. This process can be noisy: the LLM correctly flags explicit anomalies but also produces false positives on ambiguous content.

We evaluate two open-weight models from the Qwen2.5-Instruct family [Qwen Team 2025]: the 7B (≈4 GB) and 14B (≈9 GB) variants, both in 4-bit GGUF quantization, executed locally via llama.cpp [Gerganov et al. 2023]. Qwen2.5-Instruct was selected as an open-weight instruction-tuned model that combines strong multilingual instruction-following capability with a favourable performance-to-size ratio, making it well-suited for local deployment and prompt-based weak-label generation without cloud infrastructure. 4-bit quantization substantially reduces GPU memory requirements while maintaining adequate inference quality for prompt-based annotation, enabling deployment on a single consumer-grade GPU. If $|\hat{A}| \geq 8$, the encoder is fine-tuned contrastively using SetFit [Tunstall et al. 2022]; otherwise the original embeddings are kept. All documents are then re-encoded, and a semi-supervised detector is trained on the full training set using $\hat{A}$ as labeled anomaly examples. Ground-truth labels are used *only* for evaluation.

3.2. Sampling Strategies

The choice of which documents to present to the LLM matters: in a corpus where $\sim 95\%$ of instances are normal, a naive sample will contain few or no anomalies. We evaluate two strategies: **Random** uniformly samples N instances from the training pool, reflecting the true $\sim 5\%$ anomaly rate; **Coverage** partitions the embedding space into N clusters via k-means++ and selects the centroid-nearest instance from each cluster [Sener and Savarese 2018], ensuring that peripheral regions — where anomalies are more likely to reside — are represented.

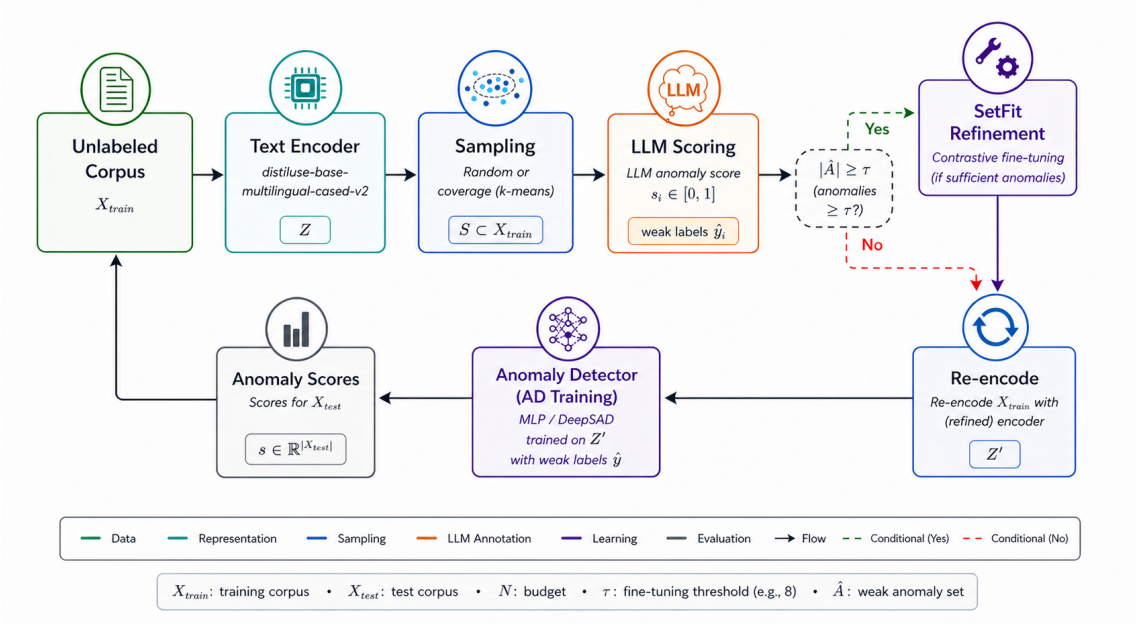


Figure 1. LLM supervision pipeline. Steps 1–3 encode the corpus, sample candidates, and query the LLM for weak anomaly labels ($\hat{A}$). Step 4 optionally refines the encoder via contrastive fine-tuning (skipped if $|\hat{A}| < 8$). Steps 5–6 re-encode and train the detector. Dashed arrow: $\hat{A}$ also provides labeled anomalies directly to the detector.

3.3. Anomaly Detection Models

We compare two representative semi-supervised AD paradigms whose training objectives respond differently to label noise.

DeepSAD [Ruff et al. 2020] learns a hypersphere in a latent space that encloses normal embeddings and pushes known anomalies away. Given encoder ϕ with centre $\mathbf{c}$, the loss for a sample (x_i, y_i) is:

$$\mathcal{L}_{\text{SAD}} = \frac{1}{n} \sum_{i=1}^{n} \begin{cases} \|\phi(x_i) - \mathbf{c}\|^2 & \text{if } y_i \in \{0, -1\} \\ \max(0,\ m - \|\phi(x_i) - \mathbf{c}\|^2) & \text{if } y_i = 1 \end{cases} \tag{3}$$

where $y_i \in \{0, 1, -1\}$ denotes normal, anomalous, or unlabeled, and m is a margin. A *single mislabeled* positive forces the encoder to push a normal instance away from the centre, directly distorting the boundary. We use the DeepOD [Xu 2022] implementation with a 3-layer MLP encoder (512→100→50→128) with ReLU activations, trained for 100 epochs.

MLP. A custom shallow binary classifier trained with standard cross-entropy:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{n} \sum_{i=1}^{n} \big[y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i) \big] \tag{4}$$

Because the loss averages over all samples — including thousands of clean negatives — a handful of mislabeled positives has limited influence on the decision boundary. Architecture: two hidden layers (128, 64) with ReLU, sigmoid output, trained with Adam

(lr $= 10^{-3}$) for 50 epochs. Both models receive the same 512-dimensional distiluse embeddings as input.

Our proposed method, **MLP+SF**, combines the MLP with SetFit-refined embeddings. To isolate the contribution of each component, Section 4.2 introduces a 2×2 ablation that independently varies the classifier (MLP vs. DeepSAD) and the embedding (original vs. SetFit fine-tuned).

3.4. Datasets and Evaluation

We evaluate on four datasets spanning two tasks and two languages. Each is prepared by retaining all normal instances, merging remaining classes into an anomaly set subsampled to $\rho \approx 5\%$, and splitting 80/20 into train and test sets.

Table 1. Dataset overview. Sizes reflect the corpus after contamination adjustment and train/test split.

Dataset	Normal class	Anomaly class	Task	Lang	Size
20 Newsgroups [Lang 1995]	rec.sport.hockey posts	all other newsgroups	TC	en	1,048
WikiNews	Politics articles	Non-politics articles	TC	pt	6,120
Hate Speech Tweets [Sharma 2018]	Non-hate speech	Hate speech tweets	HS	en	31,205
HateBR [Leite et al. 2020]	Non-offensive	Offensive comments	HS	pt	3,675

TC = Topic Classification; HS = Hate Speech detection.

The four datasets form a 2×2 grid of task and language, following the experimental design of [Maia and Costa 2025], and vary considerably in detection difficulty: 20 Newsgroups is the easiest (distinctive hockey vocabulary), the hate speech datasets are the hardest (hateful and neutral text share register and style), and WikiNews sits in between. The WikiNews corpus was extracted from Portuguese Wikimedia dumps following the methodology of Garcia et al. [Garcia et al. 2024], with articles categorised by topic section (Politics as normal class, all other sections as anomalies).

Performance is measured by ROC-AUC on the held-out test set, averaged over 3 seeds $\times$ 2 sampling strategies at $N = 200$ (6 runs per configuration). We compare against seven unsupervised detectors and the best semi-supervised model with ground-truth labels ($\sim 5\%$ of training set).

4. Experiments and Results

All experiments ran on an NVIDIA Tesla T4 GPU (Google Colab). LLM annotation used Qwen2.5-Instruct [Qwen Team 2025] 7B and 14B in 4-bit quantization via llama.cpp [Gerganov et al. 2023] at zero API cost; labels were persisted so that all ablations consume identical annotations. We evaluated at $N = 200$, the smallest budget that consistently produced enough weak positives to activate SetFit ($|\hat{A}| \geq 8$) across datasets; both sampling strategies yielded statistically indistinguishable results and are aggregated for variance reduction. [1]

[1]The source code, prompts, and experimental configuration are available at https://github.com/FabioMMaia/bootstrapping-text-anomaly-detection

4.1. RQ1: Do LLM Labels Improve Over Unsupervised Detection?

Table 2 presents seven unsupervised baselines, four LLM-supervised configurations, and the best semi-supervised model with ground-truth labels ($\sim 5\%$ of the training set).

Table 2. ROC-AUC across all methods and datasets. *Unsupervised*: no labels. *Ours*: proposed pipeline with LLM-generated weak labels and contrastive fine-tuning when $|\hat{A}| \geq 8$ (no human annotation); ‡proposed configuration. *Oracle*†: trained with ground-truth labels for $\sim 5\%$ of training set (upper bound). All rows: mean$_{\pm \text{std}}$ over 6 runs. Gap↑: fraction of the [best unsup, best oracle] interval recovered by our best variant.

	Method	20 News	WikiNews	HS Tweets	HateBR
Unsupervised	IForest [Liu et al. 2008]	$0.864_{\pm.045}$	$0.687_{\pm.027}$	$0.531_{\pm.024}$	$0.380_{\pm.032}$
	LOF [Breunig et al. 2000]	$0.856_{\pm.000}$	$0.676_{\pm.000}$	$0.498_{\pm.000}$	$0.561_{\pm.000}$
	OCSVM [Schölkopf et al. 2001]	$0.770_{\pm.000}$	$0.666_{\pm.000}$	$0.553_{\pm.000}$	$0.372_{\pm.000}$
	DeepSVDD [Ruff et al. 2018]	$0.748_{\pm.056}$	$0.620_{\pm.042}$	$0.503_{\pm.025}$	$0.539_{\pm.041}$
	AutoEncoder	$0.900_{\pm.003}$	$0.743_{\pm.011}$	$0.575_{\pm.003}$	$0.468_{\pm.004}$
	VAE	$0.920_{\pm.000}$	$0.751_{\pm.012}$	$0.526_{\pm.004}$	$0.381_{\pm.000}$
	HBOS [Goldstein and Dengel 2012]	$0.917_{\pm.000}$	$0.743_{\pm.000}$	$0.540_{\pm.000}$	$0.377_{\pm.000}$
Ours	DeepSAD (Qwen 7B)	$0.811_{\pm.104}$	$0.692_{\pm.062}$	$0.808_{\pm.032}$	$0.669_{\pm.082}$
	DeepSAD (Qwen 14B)	$0.949_{\pm.022}$	$0.785_{\pm.043}$	$0.827_{\pm.027}$	$0.629_{\pm.116}$
	MLP (Qwen 7B)	$0.763_{\pm.132}$	$0.718_{\pm.090}$	$0.861_{\pm.018}$	$\mathbf{0.742}_{\pm.087}$
	MLP (Qwen 14B)‡	$\mathbf{0.967}_{\pm.015}$	$\mathbf{0.885}_{\pm.027}$	$\mathbf{0.867}_{\pm.025}$	$0.695_{\pm.087}$
Oracle†	DeepSAD [Ruff et al. 2020]	$0.995_{\pm.004}$	$0.837_{\pm.034}$	$0.904_{\pm.020}$	$0.791_{\pm.025}$
	DevNet [Das et al. 2023]	$0.910_{\pm.133}$	$0.731_{\pm.073}$	$0.786_{\pm.062}$	$0.764_{\pm.046}$
	MLP	$0.998_{\pm.000}$	$0.936_{\pm.003}$	$0.946_{\pm.003}$	$0.886_{\pm.004}$
	XGBOD [Zhao et al. 2019]	$0.996_{\pm.000}$	$0.943_{\pm.002}$	$0.946_{\pm.000}$	$0.877_{\pm.004}$
	Gap↑	60%	70%	79%	56%

‡Proposed configuration. †Oracle: semi-supervised with ground-truth labels ($\sim 5\%$ of training set).

Unsupervised ceiling. The unsupervised block reveals a sharp difficulty divide: on 20 Newsgroups almost every method exceeds 0.85 AUC, but on the hate speech datasets the best barely exceeds chance (HateBR: 0.561, HS Tweets: 0.575). WikiNews sits between (0.751). This confirms hate speech as the regime where supervision is most needed — and least available.

LLM-supervised gains. Every LLM-supervised configuration surpasses the unsupervised ceiling on the two hate speech datasets. The proposed **MLP (Qwen 14B)**† achieves the best AUC on three of four datasets: HS Tweets rises from 0.575 to 0.867 (79% gap recovered), WikiNews from 0.751 to 0.885 (70%), and 20 Newsgroups reaches 0.967 (60%). On HateBR the 7B model outperforms the 14B (0.742, 56% gap recovered) — a cross-lingual pattern analysed in Section 4.2. Globally, mean AUC rises from 0.702 to 0.865 with zero human annotation. Despite using only $N = 200$ LLM-annotated candidates — of which only 7–16 are flagged as anomalous — the best variant falls 3–13 pp short of the upper bound, which uses 14–91$\times$ more anomaly labels on the three harder datasets.

4.2. RQ2: What Drives the Approach's Gains?

To disentangle each component's contribution, we use a 2×2 factorial design varying classifier (MLP vs. DeepSAD) and embedding (original vs. SetFit-refined); all cells share

identical LLM labels.

Model choice. MLP outperforms DeepSAD on three of four datasets for 7B (global mean +0.026) and all four for 14B (+0.056), with lower variance (std 0.015 vs. 0.022 on 20 Newsgroups, 14B). The cross-entropy loss averages over thousands of clean negatives, making MLP less sensitive to mislabeled positives than the DeepSAD hypersphere.

Embedding refinement. Figure 2 visualises the full 2×2 ablation. Contrastive fine-tuning via SetFit improves AUC on three of four datasets but can hurt when labels are systematically noisy. On the **hate speech datasets**, where both LLMs produce directionally correct labels, SetFit yields the largest and most consistent gains ($\Delta = +0.110$–0.117 on HS Tweets; +0.050–0.099 on HateBR). On **20 Newsgroups**, gains are modest ($\Delta = +0.017$ for 7B, +0.035 for 14B) despite reliable activation, because the encoder already separates the classes well. The exception is **WikiNews** with the 7B model: low-quality labels shift embeddings in the wrong direction ($\Delta = -0.069$), the only case where SetFit degrades performance. The 14B avoids this failure ($\Delta = +0.019$), confirming that label quality gates the benefit of contrastive refinement.

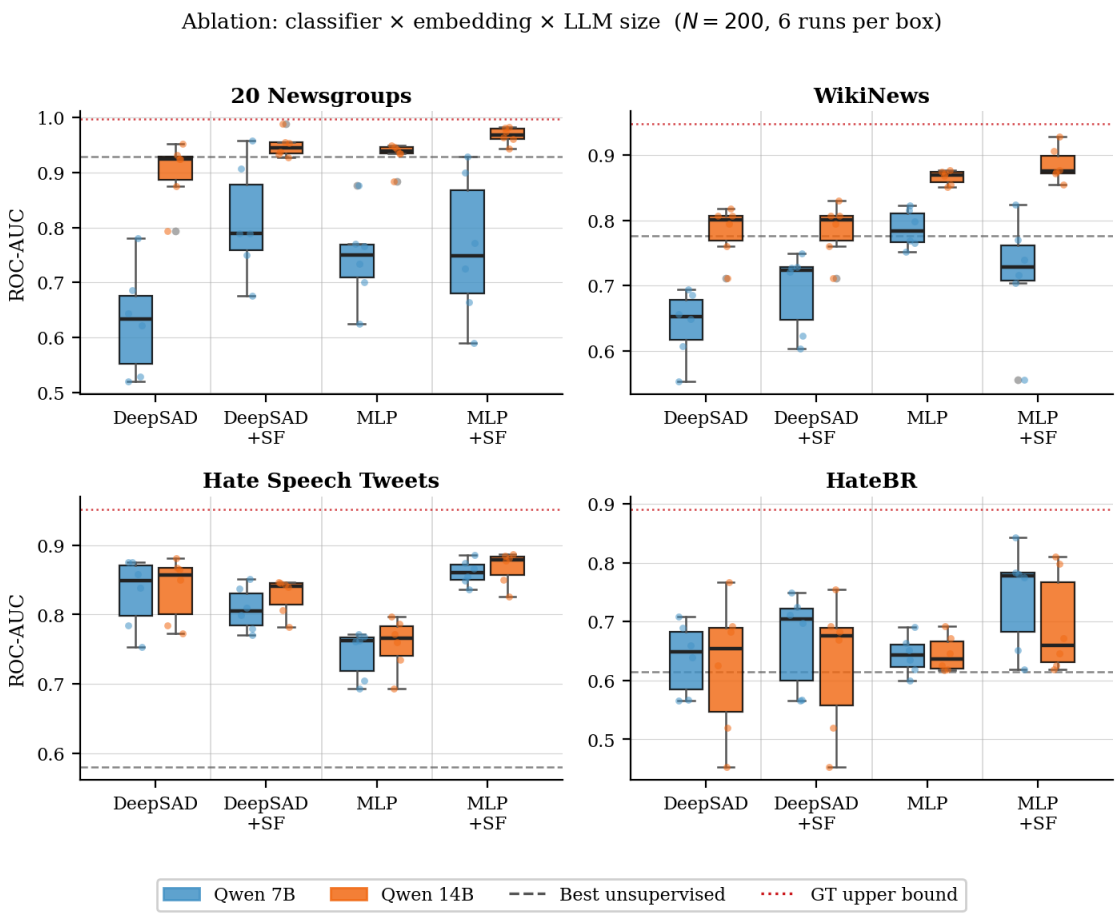


Figure 2. 2×2 **ablation at** $N = 200$ **(6 runs per box: 3 seeds** $\times$ **2 sampling strategies). Blue: Qwen 7B; orange: Qwen 14B. Dashed: best unsupervised; dotted: oracle upper bound.**

LLM capacity and noise propagation. Comparing the 14B and 7B rows in Table 2, the larger model is not uniformly superior: MLP 14B exceeds MLP 7B on WikiNews

(+0.167) and 20 Newsgroups (+0.204), but the 7B *outperforms* on HateBR (+0.047), suggesting the bottleneck for Portuguese hate speech is cultural alignment rather than capacity — validating the low-cost design (single T4 GPU). The ablation reveals a noise propagation chain: the LLM is the *source*, contrastive fine-tuning the *amplifier*, and the AD model the *filter*.

4.3. RQ3: Where Does the Approach Succeed and Fail?

The previous sections show that the approach recovers 56–79% of the gap but gains less on HateBR (14B) and 20 Newsgroups (7B). To understand *why*, we trace downstream AUC back to annotation quality (Table 3). The results reveal that annotation quality varies substantially across datasets and model sizes, yet downstream performance remains robust to low-quality labels in most cases. We identify three annotation regimes and trace their downstream impact.

Table 3. LLM annotation quality at $N = 200$ (random sampling, mean over 3 seeds). Precision and recall are computed against ground-truth labels; $\bar{n}$: mean number of positive labels returned.

Dataset	Model	Prec.	Recall	$\bar{n}$	AUC‡
20 Newsgroups	Qwen 7B	0.075	0.958	134.3	0.763
	Qwen 14B	0.172	1.000	59.3	0.967
WikiNews	Qwen 7B	0.038	0.090	24.3	0.718
	Qwen 14B	0.678	0.556	8.0	0.885
HS Tweets	Qwen 7B	0.369	0.475	13.0	0.861
	Qwen 14B	0.405	0.490	13.0	0.867
HateBR	Qwen 7B	0.565	0.463	9.7	0.742
	Qwen 14B	0.644	0.387	7.3	0.695

‡MLP downstream AUC (proposed configuration).

Implicit vocabulary gaps (20 Newsgroups). Both models achieve near-perfect recall (0.96–1.00) but low precision (0.075–0.172), flooding the weak label set with false positives from normal hockey posts that the LLM misclassifies — it fails to associate implicit domain vocabulary (team abbreviations like *Pens* for Pittsburgh Penguins, NHL jargon) with ice hockey. Crucially, the pipeline remains robust to moderate levels of this noise: the 14B still achieves 0.967 AUC because the MLP's cross-entropy loss dilutes the mislabeled positives among thousands of clean negatives, and the 14B's stronger world knowledge cuts the false-positive volume in half. The 7B's much higher noise level ($\bar{n} = 134$ flagged instances, precision 0.075) does reduce performance (AUC 0.763), marking an empirical limit of the pipeline's noise tolerance.

Criterion polarity error (WikiNews, 7B). Among all configurations, WikiNews (7B) is the only case where SetFit reduces downstream AUC relative to the corresponding non-SetFit baseline. Manual inspection of weak labels suggests that the model correctly identifies article semantics, but appears to assign the opposite label polarity for some normal political articles. For example, an article about Hugo Chávez receives score 0.8 from the 7B with the reason *"primarily discusses political actions…making it clearly political"* — a semantically correct explanation paired with a positive anomaly label. The 14B assigns the same article score 0.0, correctly recognising it as normal. This is

also the only configuration where SetFit amplifies the annotation error ($\Delta_{\text{AUC}} = -0.069$). With the 14B (precision 0.678), the same pipeline reaches AUC 0.885, suggesting that the failure originates in the annotator rather than in the downstream architecture.

Cross-lingual calibration mismatch (HateBR). On HateBR, the pipeline's downstream AUC is driven by the 7B model (0.742), which outperforms the 14B (0.695). The annotation table explains why: the 7B produces more positive labels ($\bar{n} = 9.7$ vs. 7.3) with comparable precision, suggesting greater sensitivity to low-intensity Brazilian Portuguese offense markers. The comment *"O cara não trabalha meu..."* (a personal jab) is flagged by the 7B (score 0.8) but missed by the 14B (score 0.2). The larger model appears to apply a higher decision threshold, missing culturally specific expressions and producing fewer labels that, despite higher individual precision, result in smaller downstream gains than the 7B variant.

When annotation quality is **consistent** — as on HS Tweets, where both models converge at precision ≈ 0.37–0.40 and recall ≈ 0.48–0.49 — the pipeline delivers its strongest gains (+0.110–0.117 AUC). Taken together, these regimes suggest that the pipeline is more tolerant to *quantitative* noise — large numbers of false positives — than to *directional* errors, where the annotator appears to assign the wrong label polarity. In our experiments, 20 Newsgroups (14B) reaches 0.967 AUC despite precision of only 0.172, whereas WikiNews (7B) is the only configuration where SetFit amplifies the annotation error. These observations suggest that verifying the LLM's interpretation of the normality criterion may be more important than maximising raw label count.

5. Conclusion

We presented a framework that transfers weak labels from a compact, locally deployed LLM into a lightweight text anomaly detector, requiring no human annotators, cloud APIs, or annotation budget. Across two languages and two tasks, the approach raises mean ROC-AUC from 0.702 to 0.865, closing 56–79% of the gap to oracle upper bounds.

Three findings emerge from ablating the pipeline. First, MLP consistently outperforms DeepSAD under noisy labels because cross-entropy averages over thousands of clean negatives, absorbing mislabeled positives that distort the DeepSAD hypersphere. Second, contrastive fine-tuning amplifies both good and bad label signals — yielding strong gains when annotation is reliable but degrading performance when labels are systematically inverted. Third, the pipeline tolerates high volumes of false positives(0.967 AUC despite annotation precision of 0.172) but appears less tolerant to directional noise, where the annotator may assign the wrong label polarity. This has a practical implication: practitioners should verify prompt alignment before scaling annotation.

Future work includes four directions: (i) LLM-based *data augmentation*, generating synthetic anomalies to increase the positive set beyond what the corpus naturally contains; (ii) *semantically guided sampling*, where the LLM or a richer embedding model selects candidates near the normal–anomaly boundary rather than relying on blind random or coverage strategies; (iii) *soft-label distillation*, replacing hard binary labels with continuous supervision from the LLM's calibrated probabilities to better exploit annotator uncertainty; and (iv) *deeper encoder adaptation*, such as full fine-tuning of the sentence transformer under weak supervision, to understand the trade-off between representational capacity and noise sensitivity.

6. Limitations

Our evaluation covers four datasets across two languages and two tasks, but all use a fixed $\sim 5\%$ contamination rate; performance under different rates and on other languages or domains remains untested. The LLM annotation relies on dataset-specific prompts whose sensitivity was not systematically evaluated. The approach assumes that the practitioner can articulate a normality boundary in natural language — when this prior knowledge is misspecified, annotation quality degrades (as shown in the WikiNews 7B failure mode).

7. Ethics Statement

All datasets are publicly available and previously published. The hate speech datasets contain offensive content by design; we use them solely for evaluation and do not redistribute the data. LLM annotations were generated locally without transmitting data to external APIs.

8. Generative AI Disclosure

This work uses generative AI in three roles: (i) *As the experimental method*: Qwen2.5-Instruct (7B and 14B) is used as the LLM annotator within the proposed pipeline (Section 3.2). (ii) *As a software engineering aid*: Claude Sonnet 4.6 (Anthropic) was used to assist with code prototyping, debugging, and implementation support during experimentation. (iii) *As a writing aid*: Claude Opus 4 (Anthropic) was used to assist with text revision.

All scientific content, experimental design, implementation choices, and claims are the sole responsibility of the authors.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. The authors would like to thank the National Council for Scientific and Technological Development (CNPq grant #312360/2023-1) and INCT-TILDIAR (CNPq grant #408490/2024-1).

References

Ait-Saada, M. and Nadif, M. (2023). Unsupervised anomaly detection in multi-topic short-text corpora. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Boutalbi, K., Loukil, F., Verjus, H., Telisson, D., and Salamatian, K. (2023). Machine learning for text anomaly detection: A systematic review. In *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 1319–1324.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (2000). Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, SIGMOD '00, page 93–104, New York, NY, USA. Association for Computing Machinery.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):71–97.

Das, A. S., Ajay, A., Saha, S., and Bhuyan, M. (2023). Few-shot anomaly detection in text with deviation learning. In *Neural Information Processing: 30th International Conference, ICONIP 2023, Changsha, China, November 20–23, 2023, Proceedings, Part II*, page 425–438, Berlin, Heidelberg. Springer-Verlag.

Garcia, K., Shiguihara, P., and Berton, L. (2024). Breaking news: Unveiling a new dataset for portuguese news classification and comparative analysis of approaches. *PLOS ONE*, 19(1):1–15.

Gerganov, G. et al. (2023). llama.cpp: LLM inference in C/C++. `https://github.com/ggerganov/llama.cpp`.

Goldstein, M. and Dengel, A. (2012). Histogram-based outlier score (HBOS): A fast unsupervised anomaly detection algorithm. In Wölfl, S., editor, *KI-2012: Poster and Demo Track*, pages 59–63, Saarbrücken, Germany.

He, X., Lin, Z., Gong, Y., Jin, A.-L., Zhang, H., Lin, C., Jiao, J., Yiu, S. M., Duan, N., and Chen, W. (2024). AnnoLLM: Making large language models to be better crowdsourced annotators. In Yang, Y., Davani, A., Sil, A., and Kumar, A., editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 165–190, Mexico City, Mexico. Association for Computational Linguistics.

Hendrycks, D., Mazeika, M., and Dietterich, T. (2019). Deep anomaly detection with outlier exposure. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. (2020). Supervised contrastive learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc.

Lang, K. (1995). Newsweeder: Learning to filter netnews. `http://www.cs.cmu.edu/~kenlang`. CMU, unpublished manuscript.

Leite, J. A., Silva, D. F., Bontcheva, K., and Scarton, C. (2020). Toxic language detection in social media for brazilian portuguese: New dataset and multilingual analysis. *CoRR*, abs/2010.04543.

Liu, B., Zhan, L.-M., Lu, Z., Feng, Y., Xue, L., and Wu, X.-M. (2024). How good are LLMs at out-of-distribution detection? In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8211–8222, Torino, Italia. ELRA and ICCL.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2008). Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422.

Maia, F. and Costa, A. (2025). Learning with few: A comparative study of multilingual text anomaly detection. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 233–246, Porto Alegre, RS, Brasil. SBC.

Manolache, A., Brad, F., and Burceanu, E. (2021). DATE: Detecting anomalies in text via self-supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*.

Pang, G., Shen, C., Cao, L., and Hengel, A. V. D. (2021). Deep learning for anomaly detection: A review. *ACM Comput. Surv.*, 54(2).

Qwen Team (2025). Qwen2.5 technical report. arXiv preprint arXiv:2412.15115.

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., and Ré, C. (2020). Snorkel: rapid training data creation with weak supervision. *The VLDB Journal*, 29(2-3):709–730.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S. A., Binder, A., Müller, E., and Kloft, M. (2018). Deep one-class classification. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4393–4402. PMLR.

Ruff, L., Vandermeulen, R. A., Görnitz, N., Binder, A., Müller, E., and Kloft, M. (2020). Deep semi-supervised anomaly detection. In *International Conference on Learning Representations (ICLR)*.

Ruff, L., Zemlyanskiy, Y., Vandermeulen, R., Schnake, T., and Kloft, M. (2019). Self-attentive, multi-context one-class classification for unsupervised anomaly detection on text. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4061–4071, Florence, Italy. Association for Computational Linguistics.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471.

Sener, O. and Savarese, S. (2018). Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations (ICLR)*.

Sharma, R. (2018). Twitter sentiment analysis for hate speech detection. `https://huggingface.co/datasets/tweets-hate-speech-detection/tweets_hate_speech_detection`. Accessed: 2025-09-03.

Tunstall, L., Reimers, N., Jo, U. E. S., Bates, L., Korat, D., Wasserblat, M., and Pereg, O. (2022). Efficient few-shot learning without prompts. Presented at the NeurIPS 2022 Workshop on Efficient Natural Language and Speech Processing (ENLSP-II).

Wang, S., Liu, Y., Xu, Y., Zhu, C., and Zeng, M. (2021). Want to reduce labeling cost? GPT-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xu, A., Ren, X., and Jia, R. (2023a). Contrastive novelty-augmented learning: Anticipating outliers with large language models. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11778–11801, Toronto, Canada. Association for Computational Linguistics.

Xu, H. (2022). Deepod: A benchmarking framework for deep outlier detection. `https://github.com/xuhongzuo/DeepOD`.

Xu, R. and Ding, K. (2025). Large language models for anomaly and out-of-distribution detection: A survey. In Chiruzzo, L., Ritter, A., and Wang, L., editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6007–6027, Albuquerque, New Mexico. Association for Computational Linguistics.

Xu, Y., Gábor, K., Milleret, J., and Segond, F. (2023b). Comparative analysis of anomaly detection algorithms in text data. In Mitkov, R. and Angelova, G., editors, *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1234–1245, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Yang, T., Nian, Y., Li, L., Xu, R., Li, Y., Li, J., Xiao, Z., Hu, X., Rossi, R. A., Ding, K., Hu, X., and Zhao, Y. (2025). AD-LLM: Benchmarking large language models for anomaly detection. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T., editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1524–1547, Vienna, Austria. Association for Computational Linguistics.

Zhao, Y., Nasrullah, Z., and Li, Z. (2019). Pyod: A python toolbox for scalable outlier detection. *Journal of Machine Learning Research*, 20(96):1–7.

Zhou, Z.-H. (2018). A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53.

Evaluating Portuguese Tokenizers as Morpheme Sequence in Relation to LLM Downstream Performance

Guilherme L. Mello[1,2], **Marcelo Finger**[1,2]

[1] Center for Artificial Intelligence – C4AI/USP

[2]Departamento de Ciência da Computação
Instituto de Matemática, Estatística e Ciência da Computação
Universidade de São Paulo – IME/USP

{guilhermelmello,mfinger}@ime.usp.br

Abstract. *Sub-word tokenizers allow us to handle open vocabulary problems using a relatively small set of tokens. Despite its ease of use, it usually relies on a data-driven approach that does not directly employ linguistic or morphologic features for text tokenization. [Bostrom and Durrett 2020] and [Hofmann et al. 2021] demonstrate that morphemes improve LLM performance on English texts. In this work, we explore the hypothesis that tokenizers that are capable of producing a token sequence better aligned with a morpheme sequence can improve LLMs performance on Brazilian Portuguese. To evaluate how the presence of morphemes can impact LLMs for Brazilian Portuguese, we propose MorphEval-PT, a new evaluation procedure based on the psycholinguistic concept of morphological models of word processing. We build new BPE and Unigram vocabularies that are evaluated on MorphEval-PT and, in order to validate the impact of morphemes on LLMs performance, train new LLMs from scratch and evaluate its performance on downstream tasks. Consistently, BPE demonstrates a higher precision score than Unigram in its ability to represent morphemes, as well as better performance on every downstream task. These promising results indicate that accessing the ability of tokenizers to represent morphemes is an important feature in the development of LLMs for Brazilian Portuguese and that MorphEval-PT is a good and lightweight method to improve LLM performance before any pre-training.*

1. Introduction

Different tokenization methods can be used to pre-process the input of a Large Language Models (LLMs) based on *Transformers* [Vaswani et al. 2017], ranging between word, sub-word and character level. Word level tokenization creates an extensive but closed vocabulary, resulting in several unknown words. Also, applying word level tokenizers on languages like Japanese and Chinese, when words do not have a clear boundary like white spaces, represents an extra challenge. Sub-word and character level tokenizers allow the construction of an open vocabulary, reducing the problem of Out Of Vocabulary (OOV) words by merging sub-word tokens. As a side effect of the small vocabulary, character level tokenizers create a long sequence of tokens. By including words, sub-words and characters, sub-word level tokenizers combine the flexibility of an open vocabulary with a not so long sequence of tokens. Besides its interesting features, the main sub-word level tokenizers are data-driven and create many tokens that do not represent any grammatical

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 245–257
October 19-22, 2026 ©2026 Association for Computational Linguistics

or semantical information. In any language, the meaning of words emerge from their morphemes, the smallest semantical units, and their combination follows rules of word formation [Gonçalves 2019, Plag 2018, Leminen et al. 2019]. This semantical structure will be explored in this paper.

The literature show evidences that morphology have a good impact on LLMs performance [Bostrom and Durrett 2020, Hofmann et al. 2021, Li et al. 2019, Hou et al. 2023, Park et al. 2021]. As opposed to [Bostrom and Durrett 2020], we show that BPE slightly outperform Unigram on every downstream task. We hypothesize that this gain is related with its capability of creating a more morphological coherent sequence of tokens. The main contribution of this paper are: (a) development of MorphEval-PT, a method to evaluate how likely a tokenizer can resemble a sequence of morphemes; (b) a fair comparison between Byte-Pair Encoding and Unigram tokenizers; and (c) present evidences that a morphologically coherent sequence of tokens can improve LLMs performance on downstream tasks.

In Section 2 we provide a short discussion about related works and an overview about the tokenizers algorithms and psycholinguistic models of how the brain is capable of processing the meaning of words. Section 3 describes MorphEval-PT, the evaluation framework proposed in this paper. Sections 4.1 and 4.2 describe the datasets used to train and evaluate LLMs, followed by the experiment setup in Section 5. We conclude this paper with a discussion in Section 6.

2. Related Work

Sub-word level tokenizers have become a consolidated option [Devlin et al. 2018, Radford et al. 2018, He et al. 2021, Yang et al. 2025], like WordPiece [Schuster and Nakajima 2012], Byte-Pair Encoding [Sennrich et al. 2016] e Unigram [Kudo 2018]. When introducing Unigram, [Kudo 2018] could not obtain better result than BPE on a direct comparison. But Unigram enabled the calculation of probabilities and the usage of a regularization method that improved the performance against BPE. [Bostrom and Durrett 2020] shows that BPE is a suboptimal option, assigning Unigram's better performance to its capability to create a more morphological coherent tokenization. Even though sub-word level tokenizers are efficient on generating a compact vocabulary that can handle rare and unknown words, they are still based on statistical patterns, creating several tokens that do not carry any grammatical or semantic information. In a different context, [Li et al. 2019] apply character level tokenizer on Chinese texts, a language that symbols already carry a good amount of semantic information, showing evidences that a morphological coherent sequence of tokens can improve LLMs performance.

[Hofmann et al. 2021] analyzes the influence of tokenizers on LLM's ability of interpreting complex english words. For that, they used psycholinguistic theories that try to explain how the human brain is capable of processing the meaning of complex words in a mental lexicon [Plag 2018, Leminen et al. 2019]. To improve BERT's interpretation skills, they explored derivation, a word formation process, while keeping its original vocabulary and weights. Replacing only the WordPiece algorithm by a tokenizer that generated a sequence of tokens that favored the use of the morphemes already present on the vocabulary, they allowed BERT to consume a sequence of tokens that was bet-

ter aligned with a sequence of morphemes, creating better semantic representations that improved downstream tasks performance. They also show that the same text can be understood in different ways by LLMs when considering a different sequence of tokens. [Hou et al. 2023] and [Park et al. 2021] also show the importance of considering morphology when tokenizing a text.

The main morphological resources available for Portuguese are based on Finite State Morphology [Beesley and Karttunen 2003], like Port4Nooj [Mota et al. 2016] and MorphoBr [de Alencar et al. 2018]. Even though they are linguistically precise, they are more complex and do not provide a direct sequence of morphemes. Also, they focus on inflectional morphology and do not provide enough information about derivation processes. MorphyNet [Batsuren et al. 2021] is an alternative resource that provides an easy to use dataset that contains morphological information and a morpheme segmentation extracted from Wiktionary. The evaluation of morpheme segmentation can be handled as a boundary detection problem, commonly measured by metrics like precision, recall and F1 Score [Creutz and Linden 2004, Nouri and Yangarber 2016]. We present MorphEval-PT (Section 3) by adapting MorphyNet's dataset and including the psycholinguistics concept of morphological processing in our evaluation setup.

2.1. Tokenizers

Tokenization algorithms can range between word, sub-word and character levels. Sub-word tokenizers represent a middle ground between word and character levels that combine their best features: the open and small sized vocabulary that generates a not so long sequence of tokens. In this paper we consider Unigram [Kudo 2018] and BPE [Sennrich et al. 2016] as tokenization methods. WordPiece [Schuster and Nakajima 2012] is a well known tokenization algorithm, but was never openly published by Google. Available implementations are based on public information[1]. Thus, WordPiece is usually not considered by the literature. Accordingly, we leave WordPiece out of our experiments.

BPE was introduced to solve an open vocabulary problem with a fixed sized set of sub-word tokens to significantly decrease the occurrence of OOVs. The BPE vocabulary is built in a bottom-up setup, initialized with basic symbols, and the training corpus is converted in a sequence of single character tokens. During training, the most frequent pair of tokens are iteratively merged to create a new token in the vocabulary, and the merge rule is stored in an ordered list. The training repeats until the vocabulary reaches the desired size. On text tokenization, BPE deterministically applies the merge rules in the same order as they were learned. This creates a compact sequence of tokens that proved valuable for LLMs.

Unigram follows a different strategy than BPE, building its vocabulary in a top-down approach. The vocabulary is initialized with a large set of tokens that is reduced iteratively until it reaches the required size. The vocabulary can be initialized with all possible tokens or an heuristic algorithm can be used, like Enhanced Suffix Array. At each step, the probability $p(x_i)$ is calculated for every token i. The $loss_i$ represents how much loss is introduced when the token i is removed from vocabulary. At each step, a percentage of the least meaningful tokens is removed from the vocabulary. On text

[1]WordPiece implementation: https://huggingface.co/learn/llm-course/chapter6/6

tokenization, the Viterbi algorithm can be used to find the most likely sequence of tokens x^* in the set of probable tokenizations $S(X)$.

2.2. Word Formation and The Mental Lexicon

In linguistics, morphology studies the words forms, covering the analysis of its internal structure and the processes of word formation [Gonçalves 2019, Plag 2018]. Morphemes are the smallest semantic units and, when combined, can generate words with more complex meanings. For example, in the word `morfologia`, the meaning "studies related to forms" emerges when combining the meaning of its morphemes `[morf(o)]` and `[logia]`, that mean "form" and "study", respectively. Even though morphemes are recurring parts in a language, what make them reusable is its semantic relevance [Gonçalves 2019]. Thus, words like *alho* and *bugalho*, despite having a recurring part, they do not share a morpheme as it does not have a semantic relation. The main morphological operations are derivation and inflection. Inflection creates different forms of the same lexical item, changing aspects like gender, number, tense, mood or person without changing its meaning. Derivation creates new lexical items, with new and more complex meanings. In this paper, we explore inflection and derivation by affixation because of its concatenative characteristic that resembles how sub-word tokenizers split text into tokens.

In psycholinguistics, the models of morphological process try to explain how the human brain is capable to process language and to understand the meaning of words [Plag 2018, Leminen et al. 2019]. These models contains two different routes for word processing: storage and decomposition routes. The meaning of words that are frequently used can be more efficiently recovered when stored as a single unit in the mental lexicon by the storage route. In contrast, less frequent words are more efficiently recovered when the meaning is computed by combining the meaning of its morphemes (decomposition route). [Hofmann et al. 2021] shows that it is possible to relate these mechanism with how sub-word tokenizers split text into tokens, even when tokens do not represent morphemes.

3. MorphEval-PT

Evaluating the alignment of a token sequence as a morpheme sequence is not an easy task for Portuguese, a language in which resources are limited. To partially fill this gap, we present MorphEval-PT[2], an evaluation method based on psycholinguistic models of morphological processing to access how likely sub-word tokenizers can generate a morphologically coherent sequence of tokens. MorphEval-PT is divided in two components: (i) dataset and (ii) evaluation metrics. As Portuguese does not provide the required resources to achieve the goal of this paper, we adapted MorphyNet [Batsuren et al. 2021]. Next, we introduce MorphyNet, discuss its limitations and how we handled that to build MorphEval-PT's dataset and evaluation metrics.

3.1. MorphyNet Dataset

MorphyNet [Batsuren et al. 2021] is a multilingual dataset that provides information about inflectional and derivational morphology. For inflectional morphology, it provides the lemma, inflected form, morphological features and a morpheme segmentation, as exemplified in Table 1. The derivational morphology provides the source word, target word, POS tags, a single morpheme and its type (prefix or suffix), as exemplified in Table 2.

[2]Available at: https://github.com/guilhermelmello/MorphEval-PT

Lemma	Inflected Form	Morphological Features	Morpheme Segmentation
infiltrar	infiltrar	V;1;SG;NFIN	-
infiltrar	infiltrares	V;NFIN|2;SG	infiltrar|ares
infiltrar	infiltrados	V|V.PTCP;PST|MASC;PL	infiltrar|ado|s
infiltrar	infiltradas	V|V.PTCP;PST|FEM|PL	infiltrar|ado|a|s

Table 1. Example of Portuguese inflectional information provided by MorphyNet.

Source Word	Target Word	Source POS	Target POS	Morpheme	Type
moral	amoral	J	J	a	prefix
adição	adicionar	N	V	ar	suffix
jurar	juramento	V	N	mento	suffix
haver	reaver	V	V	re	prefix

Table 2. Example of Portuguese derivational information provided by MorphyNet.

3.2. Dataset Construction and Limitations

To guarantee a better data quality, we started by removing entries where words are not present in a Brazilian Portuguese dictionary. For that, we used Enchant/Hunspell[3][4] spell checkers, that use the dictionary provided by VERO Project[5]. In this step, we kept the entries where the lemma (source word) or inflected form (target word) were validated by the spell checker. To be able to evaluate sub-word tokenizers, the MorphEval-PT dataset must contain an input word and its expected morpheme segmentation. Considering the inflectional data provided by MorphyNet, it is possible to use the inflected form as input and the morpheme segmentation as output. However, as Table 1 shows, some limitation exists: (i) only inflectional morphemes are available and (ii) the morphemes are expressed in their canonical form, which cannot recreate the inflected form only by concatenation. In respect to derivational data, the target word, morpheme and its type can be used. Also, some limitations exits: (i) only the morpheme used to create the derived word is provided by MorphyNet (see Table 2).

To obtain a sequence of morphemes that recreates the input word, different rules were applied for inflection and derivation. For inflectional data we considered the inflected form as input and transformed the sequence of morphemes from underlying to surface level. Entries that does not provide a morpheme segmentation were also removed. To create the surface level, we converted the feminine morphemes `ado|a` into `ada`. Next, we ignored the first morpheme as it represents the lemma and, using the remaining morphemes in reverse order, iteratively removed them from the inflected form. As an example, `infiltrar|ado|a|s` was converted into `infiltr|ada|s`, which concatenation results in *infiltradas* as expected. Converting derivational data was simpler and only required the removal of the affix from the derived form based on its type (prefix or suffix), resulting in a segmentation with two morphemes. For example, the words *amoral* and *adicionar* were segmented as `a|moral` and `adicion|ar`, respectively. For derivational morphology we also kept the affix type to assist during evaluation. Both derivational and inflectional data were validated to guarantee that input word could be ob-

[3] https://rrthomas.github.io/enchant/
[4] https://github.com/hunspell/hunspell
[5] https://pt-br.libreoffice.org/projetos/vero

tained by concatenating the morphemes. Any entry that did not respected it was removed from MorphEval-PT dataset. The final dataset contains 11501 derivation and 299807 inflection entries, representing 97.68% and 79.66% of MorphyNet's entries, respectively.

With MorphEval-PT's dataset in place, some usage limitations must be considered. As no data balancing were applied, some morphemes like *anti-*, *-mente* and *-s* are more frequent than others and their higher frequency can benefit tokenizers that are better at their token type. Another limitation concerns the segmentation provided by MorphyNet. As inflectional data provides only inflectional morphemes, existing derivational morphemes are not specified, like *desempates* and *prefabricados* which are segmented as `desempate|s` and `prefabric|ado|s`, respectively. For derivational data, only a single derivational morpheme is provided by MorphyNet and entries that contain more than one morpheme may appear multiple times as different entries. The word *descolonização*, for example, have different entries for each affix, providing `des|colonização` and `descoloniza|ção` as segmentations. Even though some words can be fully segmented by combining multiple entries, to evaluate the extent of this feature requires specialized knowledge and is beyond of the scope of this paper. Thus, a proper usage of this dataset is limited to affixes identifications rather than a complete morpheme segmentation.

3.3. Evaluation Metrics

To evaluate a sequence of morphemes we considered it as a boundary detection problem [Creutz and Linden 2004]. In this setup, we try to evaluate if a sequence of tokens have the same boundaries as the provided morphemes, which can be framed as a binary classification problem. For example, the targets `amoral|ismo` and `infiltr|ada|s` can be translated in the sequence of labels $[0, 0, 0, 0, 0, 1, 0, 0, 0]$ and $[0, 0, 0, 0, 0, 1, 0, 0, 0, 1]$, respectively. To handle dataset limitations, as the morphemes may not represent a complete morphological segmentation and the correctness of the base of the complex word can not be assured, we removed their boundaries from target labels. Taking the previous examples, as `amoral|ismo` and `infiltr|ada|s` does not provide the segmentation of their prefixes, we only considered `|ismo` and `|ada|s`, resulting in the label sequences $[1, 0, 0, 0]$ and $[1, 0, 0, 0, 1]$, respectively. The same is applied when the base of the word is placed on the right side of the sequence, then the segmentation `anti|moralismo` expects the sequence $[0, 0, 0, 1]$ as target labels. When relating this labels with psycholinguistics models of morphological processing, they represent the decomposition route. But sub-word tokenizers can also segment a word as a single token, which represents the storage route, have a clear semantical meaning and is also morphological coherent. To mimic both routes of the psycholinguists models this is also considered as correct segmentation by MorphEval-PT's evaluation metrics (precision, recall and F1 Score).

In the context of morphological coherent segmentation, $precision = \frac{TP}{TP+FP}$ measures how many of the boundaries created by the tokenizer were correctly placed. A higher precision indicates a higher morphological coherence while a lower precision shows that tokens are smaller units than morphemes. Two morphemes can be merged in a single token and still have a clear semantical meaning, as a morphological coherent token; thus, precision is the main metric to evaluate morphological coherent tokenization. The $recall = \frac{TP}{TP+FN}$ measures how many boundaries were correctly recovered by the tokenizer, indicating if at least the morpheme boundaries were correctly placed. As it

does not consider every boundary created by the tokenizer, character level tokenization can result in 100% recall. Therefore, recall can only show us evidences if morphemes are likely to be merged in a single token or not. The $F_1 = 2\frac{Precision \times Recall}{Precision + Recall}$ is the harmonic average between precision and recall.

4. Methods

4.1. Corpus Carolina

In order to pretrain Brazilian Portuguese LLMs from scratch we used Corpus Carolina [Crespo et al. 2023] in its 2.0.1 version, which is an open corpus with curated texts extracted from web. It includes 2108999 texts spanning different domains, like legal texts, wikis and social media. The corpus is available for download on Hugging Face[6].

4.2. Downstream Tasks

In this paper we used ASSIN [Fonseca et al. 2016], ASSIN 2 [Real et al. 2020] and HateBR [Vargas et al. 2022] to evaluate LLMs on downstream tasks. ASSIN and ASSIN 2 contain 10000 and 9448 a pair of sentences, respectively, split in training, validation and test sets that were manually annotated for Recognizing Textual Entailment (RTE) and Semantic Textual Similarity (STS). Its texts were extracted from news and contains both Brazilian and European Portuguese. STS is a regression problem and its labels range from 1 to 5, indicating the similarity level between the sentences. RTE is a classification problem with different label set for each dataset. ASSIN provide the labels *entailment*, *paraphrase* and *none*. ASSIN 2 simplified the RTE task to a binary classification by using *entailment* and *non-entailment* labels.

HateBR [Vargas et al. 2022] is an open corpus motivated by the growing hostility in social media developed to help research about offensive and hate speech detection. It contains 7000 comments extracted from social that were manually annotated in 3 levels of hostility. The fist annotation level represents the offensive language classification task, a binary classification problem that contains 3500 comments with some type of offensive language. The second annotation level if a fine-grained annotation for offensive comments (3500) that specifies a level of offensiveness: highly, moderately and slightly. The third level of annotation represents different types of hate speech present in offensive comments. In this paper we used the first level as it is and a modification of the third level to create a binary hate speech classification task. We used the splits provided in a HuggingFace dataset[7].

5. Experiments[8]

In order to evaluate the hypothesis that tokenizers capable of producing a token sequence better aligned with a morpheme sequence can improve LLMs performance on Brazilian Portuguese, we split our experiment in two steps. First, in Section 5.1, we train new tokenizers and compare them using MorphEval-PT. In the second step, Section 5.2, the tokenizers are used to pre-train new LLMs for Brazilian Portuguese and their performance are compared on downstream tasks, relating the performance on both steps.

[6]Available at: https://huggingface.co/datasets/carolina-c4ai/corpus-carolina

[7]Available at: https://huggingface.co/datasets/ruanchaves/hatebr

[8]The code used to run the experiments are available at:
https://github.com/guilhermelmello/msc_codes

5.1. Evaluating Morphological Coherence in Tokenizers

Tokenizers transform the input text into a sequence of tokens, usually, following the same pipeline: (a) normalization; (b) pre-tokenizer; (c) model; and (d) post-processing. To properly compare Unigram against BPE, we only changed the model in this pipeline. We used NFC Unicode Normalizer, which showed better results in preliminary experiments. The white space were used as pre-tokenizer. We also applied byte-level on both models to eliminate the presence of unknown tokens. In the post-processing step we introduced special tokens (`[EOS]`, `[SEP]` and `[PAD]`). We trained Unigram and BPE tokenizers using the Corpus Carolina (Section 4.1) with vocabulary sizes 5k, 8k, 10k, 30k and 50k. Tokenizers are public available on HuggingFace Hub[9].

The results for MorphEval-PT are divided by morphological process (Tables 3 and 4). In both cases, when evaluating the single token values, which represents the storage route, BPE includes more words in its vocabulary as size increases, an expected behavior of its bottom-up strategy. Even though Unigram show a similar behavior, the growth rate is considerably lower, which shows that Unigram has a preference for using the decomposition route. In our experiments BPE achieved higher precision score than Unigram, in special for inflectional morphemes with gains ranging between 10-30% over Unigram. It is worth noting that this gain is not a result of more segmentations as a single token by BPE. In Table 3, we can observe that BPE required a vocabulary of 5k to obtain a precision of 61.77% with only 0.77% of single tokens. Meanwhile, Unigram required a vocabulary of 50k to obtain a similar precision (61.07%) with 5.29% single tokens. The recall analysis require a careful inspection, as it may represent a better compression rate when associated with a high precision, which can be observed for BPE in Table 3. The opposite, high recall with low precision, is an evidence that tokens are being created unnecessarily, which can be observed for Unigram in Table 4. We also evaluate the presence of morphemes in the vocabulary of each tokenizer. BPE was also able to include more affixes in theirs vocabularies, with 20% to 30% more affixes than Unigram. This difference remained stable when increasing the vocabulary size.

When considering other models in the literature, Albertina [Santos et al. 2024] obtained the better precision and F1 scores for derivational morphemes (Table 3). This is an impressive result when considering that even though it is a Portuguese model, it used the original vocabulary from DeBERTa [He et al. 2021], which was trained on English data. For inflection data (Table 4), Qwen3 [Yang et al. 2025] also obtained impressive results for a multilingual LLM.

5.2. Impact of Morphological Coherence on Downstream Tasks

To evaluate the impact of better morphological coherent tokenizers, we trained LLMs from scratch using Unigram and BPE, called Qwen-PT-Unigram and Qwen-PT-BPE, respectively. We used the 8k vocabulary tokenizers, which show a good precision with a compact vocabulary (Section 5.1). We trained a Qwen3 [Yang et al. 2025] architecture with Causal Language Modeling (CLM) with sequences of 1024 tokens for 10 epochs on Corpus Carolina (Section 4.1). Therefore, the only difference between them is the tokenizer's model. Using batches of 1k, sentences were merged using `[SEP]` in

[9]Available at: https://hf.co/collections/guilhermelmello/tokenizers-qwen-pt

Model	Vocab		Precision	Recall	F1 Score	Single Token
BPE		5k	61.77%	45.90%	52.67%	89 (0.77%)
		8k	63.95%	43.34%	51.66%	224 (1.95%)
		10k	64.20%	42.33%	51.02%	333 (2.90%)
		15k	66.65%	41.63%	51.25%	526 (4.57%)
		30k	70.34%	41.49%	52.20%	1101 (9.57%)
		50k	75.73%	43.07%	**54.91%**	1792 (15.58%)
Unigram		5k	47.03%	68.32%	55.71%	135 (1.17%)
		8k	57.53%	66.59%	**61.73%**	195 (1.70%)
		10k	57.94%	60.86%	59.36%	225 (1.96%)
		15k	59.19%	57.62%	58.39%	291 (2.53%)
		30k	59.21%	53.59%	56.26%	465 (4.04%)
		50k	61.07%	49.80%	54.87%	608 (5.29%)
Qwen3 0.6B	BPE	150k	64.49%	46.04%	53.73%	21 (0.18%)
Albertina 900m	Unigram	128k	74.22%	55.20%	**63.32%**	106 (0.92%)
Albertina 100m	BPE	50k	46.02%	51.53%	48.62%	11 (0.10%)
PeLLE	BPE	50k	51.24%	43.27%	46.92%	68 (0.59%)
BERTimbau	WordPiece	30k	71.16%	46.89%	56.53%	861 (7.49%)
mBERT-cased	WordPiece	120k	62.72%	47.04%	53.76%	326 (2.83%)
TeenyTinyLlama	BPE	32k	70.74%	42.41%	53.03%	940 (8.17%)

Table 3. MorphEval-PT results on derivational dataset measured by precision, recall and F1 Score. Single Token represents the percentage of words that were not splitted in sub-words in a total of 11501 words, which represents the storage route.

Model	Vocab		Precision	Recall	F1 Score	Single Token
BPE		5k	51.44%	58.26%	54.64%	590 (0.20%)
		8k	53.94%	56.20%	55.04%	1207 (0.40%)
		10k	54.98%	55.38%	55.18%	1677 (0.56%)
		15k	57.25%	53.32%	55.21%	2804 (0.94%)
		30k	58.61%	51.11%	54.60%	6467 (2.16%)
		50k	62.47%	49.89%	**55.48%**	11094 (3.70%)
Unigram		5k	30.00%	52.63%	**38.22%**	926 (0.31%)
		8k	29.16%	48.36%	36.38%	1451 (0.48%)
		10k	29.25%	46.67%	35.96%	1796 (0.60%)
		15k	29.16%	44.02%	35.08%	2545 (0.85%)
		30k	30.60%	40.08%	34.71%	3792 (1.26%)
		50k	27.90%	35.00%	31.05%	5204 (1.74%)
Qwen3 0.6B	BPE	150k	58.28%	64.47%	**61.22%**	1020 (0.34%)
Albertina 900m	Unigram	128k	37.30%	43.12%	40.00%	1859 (0.62%)
Albertina 100m	BPE	50k	49.01%	65.03%	55.90%	526 (0.18%)
PeLLE	BPE	50k	26.36%	22.59%	24.33%	1903 (0.63%)
BERTimbau	WordPiece	30k	29.14%	28.22%	28.67%	5765 (1.92%)
mBERT-cased	WordPiece	120k	26.29%	25.10%	25.68%	3277 (1.09%)
TeenyTinyLlama	BPE	32k	60.76%	56.78%	58.70%	5819 (1.94%)

Table 4. MorphEval-PT results on inflectional dataset measured by precision, recall and F1 Score. Single Token represents the percentage of words that were not splitted in sub-words in a total of 299807 words, which represents the storage route.

	ASSIN		ASSIN 2		HateBR		
Model	RTE	STS	RTE	STS	Offensive	Hate	*avg*
Qwen-PT-Unigram	0.6826	0.7377	0.8537	0.7187	0.8999	0.8585	*0.7918*
Qwen-PT-BPE	0.7309	0.7454	0.8589	0.7635	0.9157	0.8871	*0.8169*
Albertina-100m	0.8081	0.8229	0.8792	0.7810	0.8905	0.8428	*0.8374*
BERTimbau-base	0.8172	0.7905	0.8991	0.7903	0.9142	0.8627	*0.8456*
BERTimbau-large	0.8254	0.6616	0.8941	0.7995	0.9249	0.8596	*0.8275*
TeenyTinyLlama-160m	0.6998	0.7110	0.8723	0.6119	0.9071	0.8142	*0.7693*
TeenyTinyLlama-460m	0.7279	0.6810	0.8799	0.6956	0.9192	0.8683	*0.7953*

Table 5. Results on ASSIN, ASSIN 2 and HateBR tasks. RTE and HateBR tasks is measured by F1 Score and STS tasks is measured by Pearson Correlation.

between and splitted in chunks of 1024 tokens. `[PAD]` tokens were inserted on left side when needed, preceded by `[EOS]` token. In preliminary experiments, smaller models obtained better performance and we used L=12, H=1024, A=16 to obtain 0.2B models.

Since it is a decoder architecture, the last token representation was used to fine-tune on ASSIN, ASSIN 2 and HateBR tasks. For each downstream task, we performed a grid search with batch size (4, 8, 16 and 32) and learning rate (5e-3, 5e-4, 5e-5 and 5e-6) for 5 epochs with 42 as a random seed. With selected hyperparameters, the fine-tuning restarted from scratch with random classification weights for 10 epochs, always selecting the best training model. Results show us that BPE consistently outperformed Unigram with an average gain of +2.51 (Table 5). In general, models based on encoder architecture obtained better results. When considering a decoder architecture, Qwen-PT models obtained better results than TeenyTinyLlama models on both ASSIN, ASSIN 2 and HateBR tasks. Also, as expected, bigger models obtained better performance.

6. Conclusions

Motivated by psycholinguistics theories of morphological processing, this paper explored the hypothesis that a sequence of tokens better aligned as a sequence of morphemes can improve LLMs performance on Brazilian Portuguese. We presented MorphEval-PT, an objective evaluation method that helps to assess how likely tokenizers can generate tokens that resemble morphemes. The promising results show evidences that collaborate with literature, but more experiments are required to confirm this hypothesis.

Even though the literature is not consistent on which tokenization method is more efficient [Bostrom and Durrett 2020, Kudo 2018], BPE consistently outperformed Unigram in our experiments, providing evidences that a morphological coherence is a good feature to assess LLM performance on downstream tasks (Section 5.2). By using MorphEval-PT metrics we could observe that BPE contained more morphemes in its vocabulary and was able to create tokens that better align with a morphological coherent segmentation (Section 5.1). Therefore, MorphEval-PT is an important tool that can be used to help the development of new LLMs for Brazilian Portuguese. By providing metrics that can be easily accessed prior any pretraining, MorphEval-PT can reduce the cost related to LLM development by assisting in iterative improvement of the tokenizer. To overcome its limitations, as future work, a more robust dataset needs to be developed.

7. Acknowledgments

This work was carried out at the Center for Artificial Intelligence of the University of São Paulo (C4AI - http://c4ai.inova.usp.br/), with support by the São Paulo Research Foundation (FAPESP 2019/07665-4) and by the IBM Corporation.

8. Limitations

As already discussed in Section 3, the proposed evaluation method MorphEval-PT contains a series of limitation inhered from MorphyNet. These limitation were addressed by the interpretation of evaluation metrics (Section 3.3). Therefore, MorphEval-PT evaluation procedure cannot state an overall capacity of tokenizers in generating a morpheme segmentation. While these limitation are not overcome, MorphEval-PT can only support evidences of the ability of generating derivational and inflectional morphemes.

As stated by [Hou et al. 2023], the gain in performance that may result by the presence of morphemes may not be statistically significant. To overcome this limitations, more experiments must be executed to support and validate if the results in this paper are significant and hold with different training settings, with different LLM architectures and vocabulary sizes. Also, other tokenizers can be considered, like Morfessor [Creutz and Lagus 2002] and StateMorph [Nouri and Yangarber 2017].

9. Disclosure on Generative AI Usage

During the development of this work, we used ChatGPT to assist with experiment's code generation. Every generated line of code was manually validated and edited by the authors. During the writing of this paper, no generative AI was used and the authors take full responsibility for the content of this publication.

References

Batsuren, K., Bella, G., and Giunchiglia, F. (2021). Morphynet: a large multilingual database of derivational and inflectional morphology. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 39–48. Association for Computational Linguistics.

Beesley, K. R. and Karttunen, L. (2003). *Finite-state morphology*. CSLI Publications.

Bostrom, K. and Durrett, G. (2020). Byte pair encoding is suboptimal for language model pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics.

Crespo, M. C. R. M., de Souza Jeannine Rocha, M. L., Sturzeneker, M. L., Serras, F. R., de Mello, G. L., Costa, A. S., Palma, M. F., Mesquita, R. M., de Paula Guets, R., da Silva, M. M., Finger, M., de Sousa, M. C. P., Namiuti, C., and do Monte, V. M. (2023). Carolina: a general corpus of contemporary brazilian portuguese with provenance, typology and versioning information.

Creutz, M. and Lagus, K. (2002). Unsupervised discovery of morphemes. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*, pages 21–30. Association for Computational Linguistics.

Creutz, M. and Linden, B. (2004). Morpheme segmentation gold standards for finnish and english.

de Alencar, L. F., Cuconato, B., and Rademaker, A. (2018). Morphobr: An open source large-coverage full-form lexicon for morphological analysis of portuguese. *Texto Livre: Linguagem e Tecnologia*, 11(3):1–25.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fonseca, E. R., dos Santos, L. B., Criscuolo, M., and Aluísio, S. M. (2016). Visao geral da avaliação de similaridade semântica e inferência textual. *Linguamática*, 8(2):3–13.

Gonçalves, C. A. (2019). *Morfologia*. Parábola, 1st edition.

He, P., Liu, X., Gao, J., and Chen, W. (2021). Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Hofmann, V., Pierrehumbert, J., and Schütze, H. (2021). Superbizarre is not superb: Derivational morphology improves BERT's interpretation of complex words. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3594–3608. Association for Computational Linguistics.

Hou, J., Katinskaia, A., Vu, A.-D., and Yangarber, R. (2023). Effects of sub-word segmentation on performance of transformer language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7413–7425. Association for Computational Linguistics.

Kudo, T. (2018). Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75. Association for Computational Linguistics.

Leminen, A., Smolka, E., Duñabeitia, J. A., and Pliatsikas, C. (2019). Morphological processing in the brain: The good (inflection), the bad (derivation) and the ugly (compounding). *Cortex*, 116:4–44.

Li, X., Meng, Y., Sun, X., Han, Q., Yuan, A., and Li, J. (2019). Is word segmentation necessary for deep learning of Chinese representations? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Mota, C., Carvalho, P., and Barreiro, A. (2016). Port4NooJ v3.0: Integrated linguistic resources for Portuguese NLP. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1264–1269. European Language Resources Association (ELRA).

Nouri, J. and Yangarber, R. (2016). A novel evaluation method for morphological segmentation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3102–3109. European Language Resources Association (ELRA).

Nouri, J. and Yangarber, R. (2017). Learning morphology of natural language as a finite-state grammar. In *Statistical Language and Speech Processing*, pages 44–57, Cham. Springer International Publishing.

Park, H. H., Zhang, K. J., Haley, C., Steimel, K., Liu, H., and Schwartz, L. (2021). Morphology matters: A multilingual language modeling analysis. *Transactions of the Association for Computational Linguistics*, 9:261–276.

Plag, I. (2018). *Word-formation in English*. Cambridge university press.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training.

Real, L., Fonseca, E., and Gonçalo Oliveira, H. (2020). The assin 2 shared task: a quick overview. In *International Conference on Computational Processing of the Portuguese Language*, pages 406–412. Springer.

Santos, R., Rodrigues, J., Gomes, L., Silva, J., Branco, A., Cardoso, H. L., Osório, T. F., and Leite, B. (2024). Fostering the ecosystem of open neural encoders for portuguese with albertina pt-* family.

Schuster, M. and Nakajima, K. (2012). Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725. Association for Computational Linguistics.

Vargas, F., Carvalho, I., Rodrigues de Góes, F., Pardo, T., and Benevenuto, F. (2022). HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7174–7183, Marseille, France. European Language Resources Association.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., and Qiu, Z. (2025). Qwen3 technical report.

Tail Smoothing and Cross-Lingual Volatility: Evaluating Estimative Uncertainty in Large Language Models for Brazilian Portuguese

Renato O. Miyaji[1], **Pedro L. P. Corrêa**[1]

[1]Escola Politécnica – Universidade de São Paulo (USP)

{re.miyaji, pedro.correa}@usp.br

Abstract. *This study evaluates how LLMs interpret Words of Estimative Probability (WEPs) in Brazilian Portuguese compared to English. We translated an English benchmark and compared multilingual models (GPT-5.1, Gemini 3 Flash) against a region-specific model (Sabiá 4). Our findings reveal a "tail smoothing" phenomenon, where models systematically compress extreme probabilities. Notably, while Gemini 3 Flash demonstrated remarkable cross-lingual stability, GPT-5.1 exhibited significant calibration degradation. Counterintuitively, when measured against the English human baseline, Sabiá 4 displayed the most aggressive distribution compression. These results suggest a complex dynamic: either linguistic fine-tuning does not ensure alignment, or it actively captures a culturally specific pragmatic ambiguity in Brazilian Portuguese. This exposes the vulnerabilities and nuances of deploying LLMs in nuanced semantic tasks in Brazilian Portuguese.*

1. Introduction

Words of estimative probability (WEPs), such as "maybe", "likely," or "probably not" are fundamental components of natural language used to communicate varying degrees of uncertainty [Erev and Cohen 1990, Juanchich and Sirota 2020]. In human communication, these linguistic markers possess complex probabilistic semantics that rely heavily on context, culture, and shared cognitive understanding [Vlasyan et al. 2018]. As artificial intelligence systems increasingly mediate critical decision-making processes, the calibration of these expressions with precise numerical estimates has become a vital area of study to ensure safe and reliable human-computer interaction [Tang et al. 2026].

The proliferation of Large Language Models (LLMs) has amplified the need for robust uncertainty communication in natural language processing. While modern LLMs generate fluent and seemingly confident text, their internal statistical representations often diverge from human cognitive interpretations of uncertainty [Schneider 2016]. Understanding whether these models assign the same probabilistic weight to WEPs as human speakers do is essential for preventing miscommunication, particularly in high-stakes domains where the precise likelihood of an event is crucial.

Recent research has highlighted significant gaps in how LLMs interpret and deploy WEPs compared to human baselines. For instance, [Tang et al. 2026] demonstrated that while established LLMs show some alignment with human estimates for specific WEPs in English, their performance degrades when confronted with cross-linguistic

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 258–269
October 19-22, 2026 ©2026 Association for Computational Linguistics

prompts or complex contextual scenarios. Their study emphasized the limits of statistical language models in reproducing the subtleties of communication under uncertainty, pointing to a pressing need for broader evaluations across diverse linguistic landscapes.

Despite these advances, the vast majority of studies on estimative uncertainty in LLMs remain heavily skewed toward English [Shorinwa et al. 2025] and, to a lesser extent, languages like Mandarin Chinese [Tang et al. 2026]. Portuguese, specifically Brazilian Portuguese, remains largely underexplored in this context, despite being a widely spoken language with its own rich set of pragmatic and semantic nuances regarding uncertainty [Freitag et al. 2022]. This gap poses a critical challenge for the equitable and accurate deployment of LLMs for Portuguese speakers, as cultural and linguistic variations can significantly alter the interpretation of probabilistic language [Schuck et al. 2025].

To address these challenges, this study is guided by two primary Research Questions: (RQ1) *"How do generalized commercial LLMs compare to language-specific models in aligning with human calibrations of WEPs in Brazilian Portuguese?"* and (RQ2) *"What are the downstream vulnerabilities introduced by misaligned uncertainty representations in this language?"*. Our main contributions include the adaptation of an English WEP benchmark into Brazilian Portuguese and a comparative analysis of general-purpose models, such as GPT-5.1 [OpenAI 2026] and Gemini 3 Flash [Google DeepMind 2026], and Sabiá 4 [Laitz et al. 2026]. By evaluating these systems, we highlight the practical limitations and inherent risks of deploying large commercial models for nuanced linguistic tasks in Portuguese. These findings are particularly critical for high-stakes applications, such as decision-making pipelines and LLM-as-a-Judge frameworks, where the subtle misinterpretation of probabilistic language can lead to skewed evaluations and unsafe outcomes. It is crucial to clarify that this study does not treat the English human baseline as the "ground truth" for Brazilian Portuguese cognition. Given that estimative uncertainty is inherently tied to cultural pragmatics, any divergence from this baseline by the models in Portuguese is analyzed through a dual lens: it may represent a breakdown in the model's cross-lingual semantic alignment, or it may indicate that the model is actively capturing a culturally distinct probabilistic mapping.

2. Related Works

Recent research on Uncertainty Quantification (UQ) in LLMs has expanded beyond traditional parametric metrics to address the unique complexities of natural language generation. While modern UQ frameworks categorize uncertainty into multiple dimensions a critical area of investigation is how models linguistically communicate this uncertainty to users [Shorinwa et al. 2025]. Studies have demonstrated that despite possessing internal statistical uncertainty, LLMs frequently generate highly decisive and persuasive responses, failing to faithfully hedge their assertions even when explicitly prompted to express doubt. This discrepancy between a model's intrinsic confidence and its linguistic decisiveness highlights a significant vulnerability in human-AI interaction, as overly confident language can easily lead to unwarranted user reliance on inaccurate outputs [Yona et al. 2024].

To systematically evaluate this communicative gap, researchers have focused on how models interpret WEPs. Empirical evaluations comparing LLM probability distributions to human baselines have revealed significant misalignments, particularly for mid-

range, context-dependent WEPs, whereas extreme terms like "almost certain" tend to show better alignment. Furthermore, a model's statistical interpretation of these words is highly sensitive to the prompt's context, with measurable distribution shifts occurring when translating between languages or when introducing gender-specific narratives [Tang et al. 2026]. Building upon these foundational findings, our work extends the evaluation of WEPs into the underexplored landscape of Brazilian Portuguese to assess whether localized fine-tuning can effectively mitigate these cross-lingual and communicative misalignments.

3. Methodology

This study's experimental design is structured to replicate and extend the foundational framework established by [Tang et al. 2026] for evaluating WEPs. Our primary objective is to systematically assess how recent advancements in LLMs impact the calibration of estimative uncertainty, particularly when transitioning from English to Brazilian Portuguese. To achieve this, we developed a bilingual experimental pipeline that mirrors the original study's conditions while introducing state-of-the-art models to the evaluation matrix to test their pragmatic alignment.

To establish a rigorous evaluation framework, our methodology leverages the dataset constructed by [Tang et al. 2026]. This dataset is built upon the empirical foundation of the WEPs originally utilized in the survey by Fagen-Ulmschneider [Fagen-Ulmschneider 2019]. Specifically, it assesses a standardized set of seventeen expressions that span the probability spectrum: *almost certain, highly likely, very good chance, probable, likely, we believe, probably, better than even, about even, we doubt, improbable, unlikely, probably not, little chance, almost no chance, highly unlikely*, and *chances are slight*. By employing these linguistic triggers, our study ensures that the evaluation captures a comprehensive range of estimative uncertainty, mirroring established baselines of human cognitive interpretation. Furthermore, the dataset rigorously tests contextual resilience through various types of structured statements designed to frame these WEPs. These include Concise Narrative Contexts, which present simple sentences to provide a direct background for the WEPs, and Extended Narrative Contexts, which introduce more prolonged scenarios with increased contextual information and clauses [Tang et al. 2026]. By adapting these diverse statement structures into Brazilian Portuguese, our study can investigate how varying levels of contextual complexity and specific narrative backgrounds influence the probabilistic estimations generated by the models in a cross-lingual setting.

The first critical phase of our methodology involved the adaptation of the dataset. To ensure rigor, we conducted a systematic translation of the original English statements and their associated WEPs into Brazilian Portuguese. This process went beyond word-for-word translation; it required careful consideration of regional nuances to preserve the ordinal probabilistic weight and contextual subtleties of each expression. Specifically, the translation was performed by two native Brazilian Portuguese speakers with proficiency in English and backgrounds in natural language processing and linguistics. The procedure involved an initial independent forward-translation phase, followed by a consensus meeting to resolve discrepancies.

For the human baseline comparison, our study utilizes the extensive em-

pirical data gathered in the original English survey by Fagen-Ulmschneider [Fagen-Ulmschneider 2019]. We explicitly note that we do not treat these English human probability judgments as the cognitive "ground truth" for Brazilian Portuguese speakers, as probabilistic intuition is heavily influenced by cultural and linguistic factors. Instead, we employ the English survey data as a reference point for cross-lingual stability. This approach allows us to measure whether models, which are predominantly trained on English corpora, maintain their internal uncertainty calibration when prompts are translated into Portuguese. This choice is necessitated by the current scarcity of established research on estimative uncertainty specifically within the Brazilian Portuguese context. While conducting a primary human survey in Portuguese is beyond the scope of this work, our methodology rigorously evaluates the semantic degradation and transferability of WEPs across languages.

To capture the current landscape of artificial intelligence, we updated the model selection to include three LLMs: GPT-5.1 [OpenAI 2026], Gemini 3 Flash [Google DeepMind 2026], and Sabiá 4 [Laitz et al. 2026]. GPT-5.1 and Gemini 3 Flash represent the frontier of generalized models, boasting massive but predominantly English-centric multilingual training corpora. In contrast, Sabiá 4 was selected as a baseline; it is a model explicitly designed and fine-tuned for Brazilian Portuguese, possessing a deeper focus on the cultural and linguistic patterns of Brazilian language. The inclusion of Sabiá 4 is central to addressing our research questions regarding regional alignment. By comparing this region-specific model against generalized ones, we aimed to evaluate whether localized training paradigms yield a higher degree of calibration with human probability estimates in Portuguese. We specifically test the hypothesis that Sabiá 4 will demonstrate less variance and better distributional accuracy when processing portuguese WEPs, effectively minimizing the semantic degradation often observed when generalized models process non-English probabilistic language.

Our prompting strategy rigorously reproduced all experimental scenarios from the original study across both languages. Every experiment was executed in both English and Portuguese. This dual-language execution allows us to isolate the specific effect of language on model performance and to determine if the commercial models exhibit altered confidence intervals or pragmatic failures when operating outside their primary linguistic domain.

During the inference phase, the models were prompted to assign precise numerical probability values to the provided statements containing the WEPs. To ensure consistency and prevent the models from generating evasive or overly verbose responses, we utilized standardized zero-shot prompting with strict output constraints. The models were instructed to return specific percentage values, which were then systematically extracted, cleaned, and aggregated to construct probability distributions for each combination of WEP, model, and language. Finally, the quantitative evaluation of these distributions focuses on measuring the statistical divergence between the LLM-generated estimates and the human survey baseline. By analyzing metrics such as median values, interquartile ranges, and overall distribution shapes, we can objectively quantify the degree of misalignment.

This experimental design not only tests the raw capabilities of modern LLMs but also rigorously exposes the potential vulnerabilities that arise when these systems are deployed for nuanced probabilistic reasoning in Brazilian Portuguese. To rigorously eva-

luate the degree of alignment between the estimates generated by the language models and human perception, we adopted an analytical approach based on robust statistical metrics, focusing on both the shape of the distributions and stochastic equality, following [Tang et al. 2026]. The first stage of our quantitative analysis utilized the Kullback-Leibler (KL) Divergence to measure the dissimilarity between the probability distributions generated by the LLMs and the baseline reference distribution obtained from the Fagen-Ulmschneider human survey [Fagen-Ulmschneider 2019]. By calculating the KL divergence for each WEP in both languages, we were able to objectively rank the models' performance. Lower KL values indicate that the model more faithfully captured the variance and concentration of human estimates, allowing us to directly compare the semantic degradation between the English tests and their counterparts translated into Brazilian Portuguese.

However, since the probability estimates provided by both humans and language models frequently exhibit asymmetrical and non-normal distributions, traditional parametric tests for comparing means would be inadequate [Tang et al. 2026]. To overcome this limitation and achieve greater statistical rigor, we employed the Brunner-Munzel test as our primary inference tool. The Brunner-Munzel test is a non-parametric method designed to compare samples from two independent populations without assuming equal variances or normal distributions, making it ideal for the stochastic nature of WEP data.

The application of the Brunner-Munzel test allowed us to evaluate the hypothesis of stochastic equality, in other words, to verify whether the probabilities assigned by an LLM tend to be systematically higher or lower than those assigned by humans. By extracting the p-values from this test, we could determine with statistical precision which models presented significant divergences from the human baseline. This cross-analysis between the KL divergence and the Brunner-Munzel test provided the empirical foundation necessary to conclude whether Sabiá 4 indeed outperforms generalist commercial models in the fine-grained calibration of uncertainties in the Portuguese language.

4. Results

Figure 1 summarizes the experimental results.

4.1. Experiments in English

The comparative analysis between the human survey baseline and the language models operating in English reveals a persistent phenomenon of "tail smoothing" across the evaluated LLMs, as presented in Table 1. When analyzing the generated probability distributions against the Fagen-Ulmschneider survey data [Fagen-Ulmschneider 2019], it becomes evident that while the models generally respect the ordinal ranking of WEPs, they often struggle to capture the extreme values of human intuition. Both KL divergence scores and Brunner-Munzel stochastic equality tests indicate that the models tend to compress their estimates toward the center, resulting in medians that are less extreme than those provided by human respondents for both very low and very high-probability expressions.

An examination of GPT-5.1's performance highlights significant misalignment, particularly in the maximum certainty zone. For high-probability WEPs such as "almost certain," human respondents anchored their estimates at a median of 95%, whereas GPT-5.1 provided a statistically distinct median of 90%. The Brunner-Munzel test confirms this divergence as highly significant ($p < 0.001$), accompanied by a pronounced

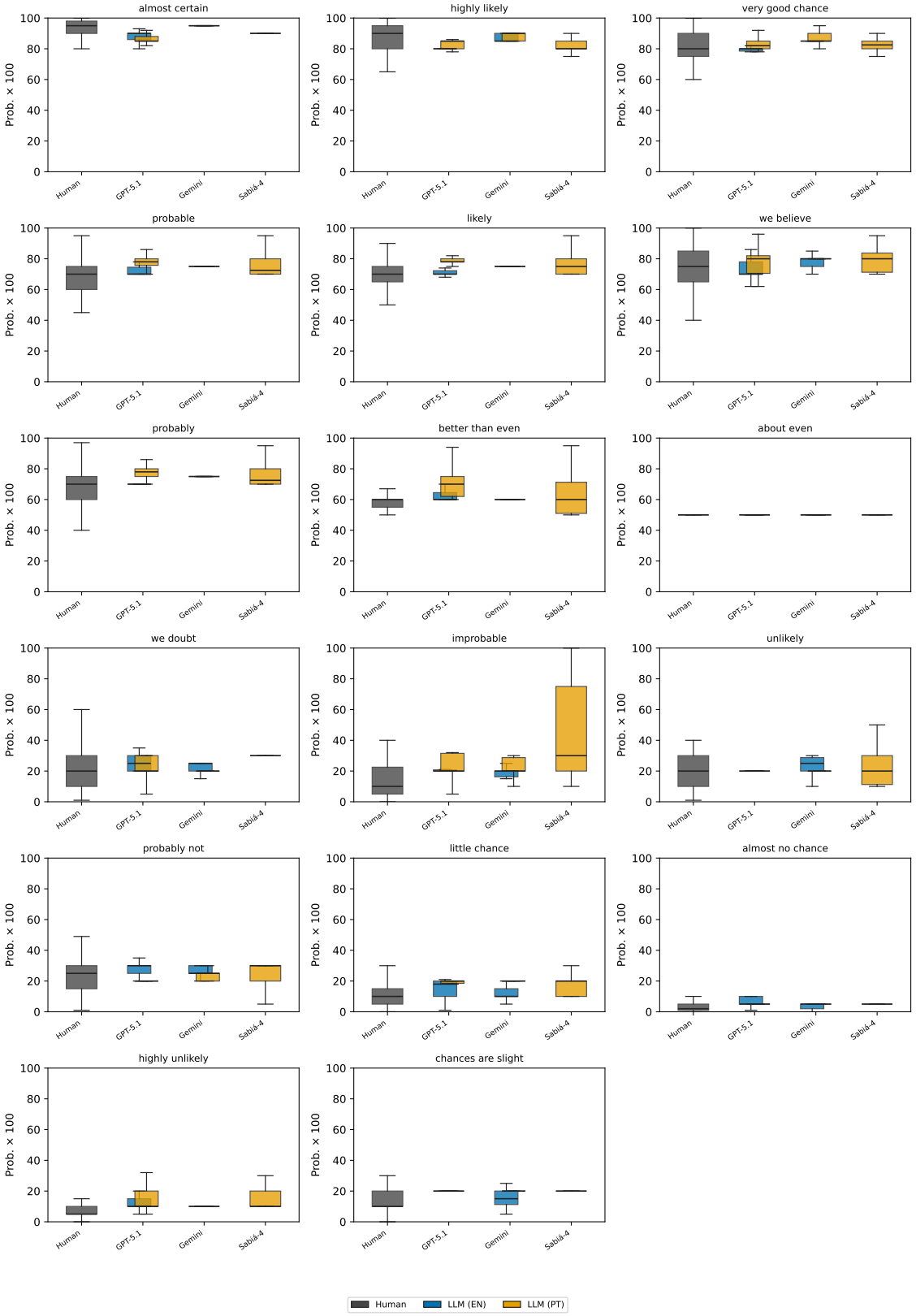


Figura 1. Comparative boxplots illustrating the probability distributions assigned to seventeen words of estimative probability. The charts contrast the Human baseline with the estimates generated by GPT-5.1, Gemini, and Sabiá-4. Model responses are distinguished by language, with LLM (EN) indicating English prompts and LLM (PT) indicating Portuguese prompts. This visualization highlights cross-lingual distribution shifts and demonstrates the ”tail smoothing”effect, where model estimations frequently compress toward the center compared to the broader variance of human intuition.

Tabela 1. English baseline: human medians vs. GPT-5.1 and Gemini 3 Flash. p_{BM}**: Brunner–Munzel test (human vs. model); KL:** $\mathrm{KL}(P_{\mathrm{human}} \| P_{\mathrm{model}})$**; AMD: absolute median difference (percentage points).**

Expression	M_{hum}	M_{GPT}	M_{Gem}	$p_{\mathrm{BM}}^{\mathrm{GPT}}$	$p_{\mathrm{BM}}^{\mathrm{Gem}}$	$\mathrm{KL}^{\mathrm{GPT}}$	$\mathrm{KL}^{\mathrm{Gem}}$	$\mathrm{AMD}^{\mathrm{GPT}}$	$\mathrm{AMD}^{\mathrm{Gem}}$
Almost certain	95	90	95	< 0.001	0.537	14.98	6.65	5	0
Highly likely	90	80	85	< 0.001	0.061	12.90	11.59	10	5
Very good chance	80	80	85	0.014	0.024	2.31	5.21	0	5
Probable	70	70	75	0.300	< 0.001	4.81	10.05	0	5
Likely	70	70	75	0.247	0.012	3.46	7.52	0	5
We believe	75	70	80	0.014	0.116	6.27	3.30	5	5
Probably	70	70	75	0.789	0.002	6.28	10.11	0	5
Better than even	60	60	60	< 0.001	0.002	8.74	10.25	0	0
About even	50	50	50	0.971	0.154	1.19	1.67	0	0
We doubt	20	25	25	0.002	0.342	9.76	9.59	5	5
Improbable	10	20	20	< 0.001	0.002	4.70	8.00	10	10
Unlikely	20	20	25	0.906	0.010	8.99	5.78	0	5
Probably not	25	30	25	0.105	0.596	5.09	7.00	5	0
Little chance	10	18	10	< 0.001	0.037	5.73	2.92	8	0
Almost no chance	2	5	5	< 0.001	0.003	1.76	3.26	3	3
Highly unlikely	5	10	10	< 0.001	< 0.001	1.80	3.64	5	5
Chances are slight	10	20	15	< 0.001	< 0.001	14.33	3.72	10	5

KL divergence score. This underestimation trend continues with the term "highly likely," where the human median of 90% is sharply contrasted by the model's median of 80% ($p < 0.001$). These metrics strongly suggest that GPT-5.1 systematically hesitates to assign the near-absolute probabilistic weight that English speakers naturally attribute to these specific terms.

In contrast, Gemini 3 Flash demonstrates a higher degree of stability and closer alignment with human baselines at the upper extremes of the probability spectrum. For the term "almost certain," Gemini successfully matched the human median of 95%. Furthermore, the Brunner-Munzel test yielded a p-value of 0.537, indicating no statistically significant difference between the human and model distributions, which is also reflected in a substantially lower KL divergence score compared to GPT-5.1. Gemini also showed borderline alignment for "highly likely," achieving a median of 85% with a non-significant divergence ($p = 0.06$), suggesting that its internal representation of high-confidence English WEPs is more robustly calibrated to human expectations.

Despite Gemini's success at the upper tail, both models exhibited shared vulnerabilities when processing low-probability WEPs. For terms like "we doubt" and "improbable," both GPT-5.1 and Gemini 3 Flash generated estimates that were systematically higher than human baselines. Humans assigned medians of 20% and 15% to these terms, respectively. In contrast, both LLMs shifted these estimates upwards, returning medians of 25% for "we doubt" and 20% for "improbable." The Brunner-Munzel tests confirm that these upward shifts are statistically significant across both models. This indicates a shared architectural or training-induced reluctance to generate extremely low probability percentages, effectively truncating the lower tail of the uncertainty distribution.

This baseline English analysis establishes that AMD alone are insufficient for capturing the complex behavior of LLMs. The combined application of KL divergence and

non-parametric testing proves that even when models closely approximate human medians, the underlying distributional shapes can fundamentally differ. While models like Gemini 3 Flash show promising calibration at the high end of the spectrum, the persistent tail-smoothing effects across commercial models in English.

4.2. Experiments in Brazilian Portuguese

The transition from English to Brazilian Portuguese reveals complex shifts in how Large Language Models handle estimative uncertainty, demonstrating that the "tail smoothing" phenomenon persists but manifests differently depending on the model, as presented in Table 2. When comparing the Portuguese WEP estimates to the human survey baseline, it becomes evident that the models still tend to compress probabilities toward the center of the distribution. However, the degree of semantic degradation and the alignment with human expectations vary among the evaluated models, exposing distinct vulnerabilities.

Tabela 2. Portuguese: human respondent medians vs. GPT-5.1, Gemini 3 Flash, and Sabiá 4. Subscripts: G = GPT-5.1, Gm = Gemini, S = Sabia 4.

Expression	M_{hum}	M_{G}	M_{Gm}	M_{S}	p_{G}	p_{Gm}	p_{S}	KL_{G}	$\mathrm{KL}_{\mathrm{Gm}}$	KL_{S}	$\mathrm{AMD}_{\mathrm{G}}$	$\mathrm{AMD}_{\mathrm{Gm}}$	$\mathrm{AMD}_{\mathrm{S}}$
Quase certo	95	85	95	90	< 0.001	0.274	< 0.001	2.54	2.96	1.06	10	0	5
Muito provável	90	85	90	80	< 0.001	0.455	< 0.001	12.07	11.56	2.19	5	0	10
Grande chance	80	82	85	82.5	0.629	< 0.001	0.573	1.54	5.83	1.77	2	5	2.5
Bem provável	70	78	75	75	< 0.001	< 0.001	0.028	4.38	11.60	6.36	8	5	5
Provável	70	78	75	72.5	< 0.001	< 0.001	0.004	4.47	13.97	8.19	8	5	2.5
Acreditamos	75	80	80	80	0.299	0.002	0.380	3.21	6.15	4.67	5	5	5
Provavelmente	70	78	75	72.5	< 0.001	< 0.001	0.005	5.92	14.01	8.33	8	5	2.5
Mais que provável	60	70	60	60	< 0.001	< 0.001	0.859	9.12	9.88	6.05	10	0	0
Cerca de 50/50	50	50	50	50	0.956	0.056	0.334	1.35	1.62	1.64	0	0	0
Duvidamos	20	20	20	30	0.015	0.402	< 0.001	6.84	9.33	10.13	0	0	10
Improvável	10	20	20	30	< 0.001	< 0.001	< 0.001	8.41	7.67	11.00	10	10	20
Pouco provável	20	20	20	20	0.397	0.550	0.351	8.62	13.05	8.04	0	0	0
Provavelmente não	25	20	25	30	0.146	0.379	0.021	8.04	4.92	6.83	5	0	5
Pouca chance	10	20	20	20	< 0.001	< 0.001	< 0.001	3.73	6.45	7.43	10	10	10
Quase nenhuma chance	2	5	5	5	< 0.001	< 0.001	< 0.001	2.14	3.29	13.06	3	3	3
Muito improvável	5	10	10	10	< 0.001	< 0.001	< 0.001	6.09	3.97	13.59	5	5	5
Chances remotas	10	20	20	20	< 0.001	< 0.001	< 0.001	11.88	3.07	6.80	10	10	10

An analysis of GPT-5.1 reveals a degradation in calibration at the high-certainty spectrum when operating in Portuguese. For the Portuguese translation of "almost certain," human respondents anchored their estimates at a median of 95%. In contrast, GPT-5.1 dropped to a median of 85%, resulting in an AMD of 10.0 and a highly significant Brunner-Munzel p-value ($p < 0.001$). A similar significant divergence occurred for "highly likely" ($p < 0.001$). This indicates a cross-lingual vulnerability; while GPT-5.1 hesitated in English, its reluctance to assign high probabilistic weight is amplified in Portuguese. Conversely, Gemini 3 Flash demonstrates exceptional cross-lingual robustness and stability, acting as the most reliable proxy for human probabilistic intuition in this language. For high-probability expressions like "almost certain" and "highly likely," Gemini perfectly matched the human medians of 95% and 90%, respectively. The Brunner-Munzel tests yielded non-significant p-values, paired with low KL divergence scores. While Gemini still exhibits minor tail-smoothing at the absolute lower end—shifting the median of "improbable" from the human 15% to 20% ($p < 0.001$)—it avoids the severe central compression seen in other models.

The performance of Sabiá 4 presents a counter-intuitive finding. Despite being explicitly fine-tuned for Brazilian Portuguese, Sabiá 4 exhibited the most aggressive tail-

smoothing. It failed to capture the extreme distributions of human intuition, assigning a median of 90% to "almost certain" and compressing low-probability terms. For instance, Sabiá 4 assigned a median of 30% to the Portuguese equivalent of "we doubt". The Brunner-Munzel tests confirmed these misalignments as significant, indicating that localized linguistic training did not translate into accurate probabilistic calibration for this model. The analysis underscores that linguistic localization does not automatically guarantee adherence to established English baselines for estimative uncertainty.

The contrast between Gemini 3 Flash and Sabiá 4 highlights the complex nature of uncertainty representation in LLMs, raising the question of whether compressed distributions reflect calibration failures or localized linguistic nuances. Regardless, for high-stakes applications in Brazilian Portuguese, these results expose the practical risks of assuming semantic parity across cultures and commercial LLM outputs.

4.3. Mutilingual Comparison

A direct comparative analysis of model performance across the language barrier reveals distinct sensitivities to linguistic framing, as presented in Table 3. By isolating the language variable and comparing the English and Portuguese output distributions directly, we can assess whether these models possess a language-agnostic internal representation of uncertainty, or if their probabilistic calibrations are tightly bound to the specific language of the prompt. The data supports the latter, demonstrating that translation induces measurable shifts in probability estimation.

Tabela 3. English vs. Portuguese comparison in responses from the *same* model (GPT-5.1 and Gemini 3 Flash). p_{BM}: Brunner–Munzel test between EN and PT distributions; KL: $\mathrm{KL}(P_{\mathrm{EN}} \| P_{\mathrm{PT}})$ (`kl_en_vs_pt` in the data); AMD: absolute EN–PT median difference (percentage points).

Expression	$\mathrm{KL}^{\mathrm{GPT}}$	$\mathrm{KL}^{\mathrm{Gem}}$	$p_{\mathrm{BM}}^{\mathrm{GPT}}$	$p_{\mathrm{BM}}^{\mathrm{Gem}}$	$\mathrm{AMD}^{\mathrm{GPT}}$	$\mathrm{AMD}^{\mathrm{Gem}}$
Almost certain	2.36	0.04	0.002	0.337	5	0
Highly likely	1.73	0.10	< 0.001	0.062	5	5
Very good chance	1.85	0.70	0.001	0.001	2	0
Probable	2.61	1.77	< 0.001	0.055	8	0
Likely	4.21	0.98	< 0.001	0.025	8	0
We believe	2.88	0.65	< 0.001	0.036	10	0
Probably	2.87	1.72	< 0.001	0.188	8	0
Better than even	1.88	1.32	< 0.001	0.443	10	0
About even	3.42	0.40	0.937	1.000	0	0
We doubt	3.16	10.02	0.055	0.034	5	5
Improbable	2.14	1.50	0.454	< 0.001	0	0
Unlikely	4.01	14.43	0.715	0.002	0	5
Probably not	3.68	1.81	< 0.001	0.004	10	0
Little chance	1.89	3.05	0.039	< 0.001	2	10
Almost no chance	0.92	0.61	0.431	0.029	0	0
Highly unlikely	1.47	0.91	0.044	0.283	0	0
Chances are slight	1.37	2.31	0.348	< 0.001	0	5

GPT-5.1 exhibits significant cross-lingual instability, demonstrating statistically significant shifts in its probability distributions when processing Portuguese compared to English. For instance, in the high-certainty spectrum, the transition from "almost certain" to its Portuguese equivalent resulted in an Absolute Median Difference (AMD) of 5.0. More importantly, the Brunner-Munzel test confirms this shift is not mere noise.

A similar degradation occurs with "highly likely," which also displays an AMD of 5.0 and a significant p-value between its English and Portuguese inferences. This indicates that GPT-5.1 does not map these expressions symmetrically; its probabilistic confidence inherently drops when the semantic environment switches to Portuguese, exposing a vulnerability in its multilingual alignment.

In contrast, Gemini 3 Flash demonstrates remarkable cross-lingual stability, acting as a robust multilingual reasoner. For the vast majority of WEPs, Gemini's English and Portuguese distributions are statistically indistinguishable. Looking at "almost certain," Gemini maintained a perfect cross-lingual median alignment, with a Brunner-Munzel p-value of 0.336, indicating no significant difference between how it handles the concept in either language. Mid-range terms like "about even" were also perfectly mirrored. While minor discrepancies appeared in the lower tail, Gemini largely preserved the semantic equivalence of estimative uncertainty across the linguistic divide.

These comparisons are crucial. They prove that simply translating a prompt does not guarantee equivalent cognitive or statistical processing by an LLM. The pronounced linguistic effects observed in GPT-5.1 suggest that its English-centric training data dictates its probabilistic confidence, which fails to cleanly project onto Brazilian Portuguese. Conversely, Gemini's performance provides empirical evidence that it is feasible to build multimodal systems capable of maintaining stable and language-agnostic representations of complex pragmatic concepts like uncertainty.

For practitioners and researchers, particularly those deploying LLMs for Brazilian Portuguese, these findings carry practical implications. It becomes evident that prompt engineering and probability calibration are not perfectly transferable across languages. Developers relying on generalized models like GPT-5.1 cannot assume that an English prompt carefully tuned for a specific probability threshold will yield the same probabilistic reasoning when localized. This cross-lingual volatility underscores the necessity of language-specific validation phases, especially in high-stakes applications like automated decision-making.

5. Conclusions

This study demonstrates that while modern LLMs exhibit high linguistic fluency, their internal representations of estimative uncertainty remain significantly misaligned with human cognitive intuition. Our comparative analysis reveals a persistent "tail smoothing" phenomenon, where models systematically avoid extreme probabilistic values. Notably, while Gemini 3 Flash showed cross-lingual stability, the region-specific Sabiá 4 exhibited aggressive distribution compression. This suggests that localized fine-tuning may capture distinct cultural pragmatics rather than simply failing to align with English-centric probabilistic baselines.

Future research should prioritize the development of a Brazilian Portuguese human baseline to ensure that evaluation benchmarks are more culturally grounded than current English-centric models. Furthermore, it is essential to explore how alternative prompting strategies and calibration techniques might mitigate the cross-lingual volatility. Extending this methodology to a broader range of markers and diverse linguistic contexts will be crucial for developing language-agnostic and reliable AI systems.

Limitations

Our study presents important limitations regarding its experimental design and evaluation framework. Primarily, our baseline for human cognitive interpretation relies on empirical data gathered in an English survey [Fagen-Ulmschneider 2019]. We recognize that utilizing English data as a baseline for Portuguese inferences is a theoretical limitation, as it does not account for the native pragmatic interpretation of Brazilian speakers. We adopt this approach to measure the cross-lingual stability of the models rather than making absolute claims about human-model alignment in Brazil. Without a localized human baseline, it remains theoretically challenging to definitively disentangle whether the distributional shifts observed (such as Sabiá 4's aggressive tail-smoothing) represent a mathematical failure in model calibration or an accurate capture of culturally distinct Brazilian Portuguese pragmatics. Conducting a primary human survey in Portuguese that matches the rigorous quality standards and demographic scale of the existing baseline involves prohibitive operational costs at this stage. Consequently, developing a native Brazilian Portuguese human baseline remains an urgent priority for future work to ensure that evaluation benchmarks are truly culturally grounded rather than English-centric.

Ethics Statement

From an ethical perspective, this research highlights vulnerabilities in human-AI interaction, particularly concerning the equitable and accurate deployment of LLMs for non-English speakers. The vast majority of studies on estimative uncertainty remain heavily skewed toward English, leaving widely spoken languages like Brazilian Portuguese largely underexplored. This gap poses a significant challenge, as cultural and linguistic variations can significantly alter the interpretation of probabilistic language. As LLM-based systems increasingly mediate critical decision-making processes, ensuring the safe and reliable communication of uncertainty is vital. Our research emphasizes that overly confident language generated by models can easily lead to unwarranted user reliance on inaccurate outputs.

Generative AI Usage in Paper Preparation

The use of Generative AI tools in the preparation of this manuscript was strictly limited to language editing and refining the academic writing.

Referências

Erev, I. and Cohen, B. L. (1990). Verbal versus numerical probabilities: efficiency, biases, and the preference paradox. *Organizational Behavior and Human Decision Processes*, 45:1–18.

Fagen-Ulmschneider, W. (2019). Perception of probability words. University of Illinois at Urbana-Champaign.

Freitag, R. M. K., Cardoso, P. B., and Tejada, J. (2022). Linguistic and paralinguistic constraints on the function of (eu) acho que as discourse markers in brazilian portuguese: A multilevel approach. *Pragmatics & Cognition*, 29(2):324–346.

Google DeepMind (2026). Gemini 3 flash. Technical report. Multimodal large language model developed by Google DeepMind.

Juanchich, M. and Sirota, M. (2020). Do people really prefer verbal probabilities? *Psychological Research*, 84:2325–2338.

Laitz, T., Almeida, T. S., Abonizio, H., Malaquias Junior, R., Bonás, G. K., Piau, M., Larcher, C., Pires, R., and Nogueira, R. (2026). Sabiá-4 technical report. *arXiv preprint arXiv:2603.10213*.

OpenAI (2026). Gpt-5.1. Technical report. Large language model developed by OpenAI.

Schneider, S. (2016). Communicating uncertainty: A challenge for science communication. In Drake, J., Kontar, Y., Eichelberger, J., Rupp, T., and Taylor, K., editors, *Communicating Climate-Change and Natural Hazard Risk and Cultivating Resilience*, volume 45 of *Advances in Natural and Technological Hazards Research*. Springer, Cham.

Schuck, A. d. F., Garcia, G. L., Manesco, J. R. R., Paiola, P. H., et al. (2025). Evaluating large language models for brazilian portuguese sentiment analysis: A comparative study of multilingual state-of-the-art vs. brazilian portuguese fine-tuned llms. *Journal of the Brazilian Computer Society*, 31(1):885–917.

Shorinwa, O., Mei, Z., Lidard, J., Ren, A. Z., and Majumdar, A. (2025). A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Computing Surveys*, 58(3).

Tang, Z., Shen, K., and Kejriwal, M. (2026). An evaluation of estimative uncertainty in large language models. *npj Complex*, 3:8.

Vlasyan, G. R. et al. (2018). Linguistic hedging in the light of politeness theory. In *European Proceedings of Social and Behavioural Sciences*. Future Academy.

Yona, G., Aharoni, R., and Geva, M. (2024). Can large language models faithfully express their intrinsic uncertainty in words? In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7752–7764, Miami, Florida, USA. Association for Computational Linguistics.

Textbook-Enriched Training for Language Models: Boosting Answer Quality in Specialized Contexts

Lucas B. Bulcão Mota[1], Larrissa Dantas[1], Daniela Barreiro Claro[1], Aline Paes[2], Claudia Freitas[1,3], Marlo Souza[1], Helena Caseli[4], Livy Real[5,6]

[1]FORMAS Research Center on Data and Natural Language
Institute of Computing, Federal University of Bahia (UFBA)
Salvador, BA, Brazil
{lucasbulcao,larrissasilva,dclaro,msouza1}@ufba.br
[2]Institute of Computing, Fluminense Federal University (UFF)
Niterói, RJ, Brazil
alinepaes@ic.uff.br
[3]Institute of Letters, Federal University of Bahia (UFBA)
Salvador, BA, Brazil
maclaudia.freitas@gmail.com
[4]Federal University of São Carlos (UFSCar), São Carlos, SP, Brazil
helenacaseli@ufscar.br
[5]Kunumi Institute, Belo Horizonte, MG, Brasil
[6]Federal University of Amazonas, Manaus, AM, Brasil.
livy@kunumi.com

Abstract. *The use of textbooks as primary sources of information has increasingly given way to tools based on Large Language Models (LLMs), raising concerns about the reliability of generated answers. This study investigates how different adaptation strategies shape the behavior of small language models in educational Question Answering (QA) tasks in Portuguese. To support this analysis, we built a question-answer dataset derived from an NLP textbook and compared base models and the Retrieval-Augmented Generation (RAG) pipeline with models adapted through supervised fine-tuning and Continued Pretraining. The evaluation relies on questions from the LARI dataset, which has been validated by human specialists, and combines automatic and qualitative assessment procedures. The findings indicate that small models tuned with structured instructional knowledge achieve stronger semantic alignment and produce more pertinent answers in Portuguese educational QA scenarios.*

1. Introduction

Large Language Models (LLMs) have increasingly been incorporated into educational settings as tools for information retrieval and knowledge generation [Yan et al. 2024, Xing et al. 2025]. Although these models can produce fluent and context-aware responses, recent studies report limitations in factuality, *grounding*, and the reliability of generated content, particularly in technical domains [Ji et al. 2023, Huang et al. 2025]. In *Question Answering* (QA) tasks, such limitations become especially problematic, since conceptual inconsistencies may directly affect the learning process.

Strategies such as *Supervised Fine-Tuning* (SFT), aimed at model specialization, and *Retrieval-Augmented Generation* (RAG), which incorporates external knowledge during generation, have been employed to enhance the performance of LLMs in specialized

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 270–284
October 19-22, 2026 ©2026 Association for Computational Linguistics

domains [Lewis et al. 2020, Gao et al. 2023]. Even so, the behavior of small language models under these adaptation settings in Portuguese educational scenarios remains insufficiently studied.

At the same time, technical textbooks have been used as sources of specialized knowledge for language models [Gunasekar et al. 2023]. In the context of Natural Language Processing (NLP) in Portuguese, the book *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português* [Caseli and Nunes 2026], developed by the Brasileiras em PLN (BPLN) group, stands out as one of the main reference materials in the field. Alongside this initiative, the LARI *dataset* [Junqueira et al. 2026] was built from the BPLN NLP textbook and contains context-question-answer triples evaluated by human specialists.

In this work, we rely exclusively on the third edition of the textbook [Caseli and Nunes 2024] as the knowledge source for both the construction of an instruction-based question-answer pairs dataset, named here as *Carapicu-textbook*, and the adaptation of small language models to act as experts into the book. We investigate two adaptation strategies: supervised fine-tuning (SFT) using *Carapicu-textbook* and continued pretraining (CPT) on the raw content of the book followed by SFT (CPT-SFT). The experiments include base models of different scales, adapted variants of these models, and a RAG pipeline built from the textbook content.

Although recent studies have examined LLM adaptation and contextual retrieval for QA in Portuguese [Costa 2024], analyses involving small models adapted with structured instructional knowledge and evaluated on educational *benchmarks* validated by human specialists are still scarce. The results indicate that adapted small models can outperform larger non-specialized models in semantic alignment and answer relevance metrics.

2. Related Work

LLM adaptation strategies have been widely studied as a way to align pretrained models with specialized domains. Among these approaches, *Supervised Fine-Tuning* (SFT) incorporates domain knowledge directly into model parameters [Ouyang et al. 2022, Gururangan et al. 2020] through instruction-based datasets. Research on parametric knowledge indicates that LLMs can store factual information within their parameters [Petroni et al. 2019, Roberts et al. 2020]. Even so, gains in automatic evaluation metrics do not necessarily correspond to higher factual fidelity or semantic consistency [Ji et al. 2023]. Models adapted exclusively through parametric learning also remain prone to generating incorrect information, a limitation that becomes particularly problematic in *Question Answering* (QA) tasks.

As an alternative, *Retrieval-Augmented Generation* (RAG) approaches integrate external retrieval mechanisms into the text generation process [Lewis et al. 2020]. Recent studies report improvements in factuality and *grounding* when retrieval-based strategies are adopted [Gao et al. 2023, Mallen et al. 2023]. At the same time, these methods still face issues related to irrelevant document retrieval and the handling of long contextual inputs [Liu et al. 2024, Asai et al. 2024].

Within the Brazilian context, recent studies have started to examine the behavior of LLMs in Portuguese for tasks such as summarization and QA [Assis et al. 2025].

At the same time, the growing availability of small *open-weight* models has intensified discussions around computational cost and model adaptation in resource-constrained environments.

QA *benchmarks* play a central role in the evaluation of language models. Resources such as SQuAD [Rajpurkar et al. 2016] helped establish this evaluation paradigm, although they were originally designed for English. In this setting, the LARI *dataset* [Junqueira et al. 2026] introduces a Portuguese educational QA resource derived directly from the BPLN NLP textbook [Caseli and Nunes 2024]. The dataset consists of context–question–answer triples reviewed by human specialists and was designed specifically for educational scenarios in Brazil.

Despite recent progress in LLM adaptation, contextual retrieval, and QA in Portuguese [Costa 2024], studies examining small models adapted with structured instructional knowledge and evaluated on educational *benchmarks* validated by human specialists remain scarce. The influence of controlled didactic materials on the textual and conceptual behavior of adapted models for educational QA in Portuguese also remains largely unexplored.

3. Resources and Datasets

This section presents the main resources used in the experiments, including the textbook adopted as the knowledge source (Section 3.1), the evaluation *dataset* (Section 3.2), and the models considered in the study (Section 3.3).

3.1. Textbook: BPLN book

As the primary knowledge source for model adaptation and evaluation, we used the third edition of the book *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*[1] [Caseli and Nunes 2024], hereafter referred to as the BPLN NLP textbook. The book covers theoretical foundations, linguistic resources, and contemporary NLP applications focused on Brazilian Portuguese, and has become one of the main educational references in the field.

The choice of this textbook was motivated by the presence of structured technical content reviewed by specialists and closely aligned with NLP education in Portuguese. Unlike heterogeneous corpora automatically collected from the Web, the textbook presents greater terminological and conceptual consistency across chapters and topics. Table 1 reports general statistics of the book computed using the Qwen3 0.6B tokenizer.

In the context of this study, the content of the BPLN NLP textbook was used as the sole knowledge source for both parametric model adaptation and the document collection employed in the RAG-based *pipeline*. The textual content of the book was organized at the chapter and section levels while preserving its original thematic structure. Details regarding text processing and the use of this resource in each experimental strategy are presented in Section 4.

[1] `https://brasileiraspln.com/livro-pln/3a-edicao/`

Table 1. General statistics of the textbook computed using the Qwen3 0.6B tokenizer

Metric	Value
Total number of chapters	33
Total number of sections	125
Total number of textbook tokens	126103
Average number of tokens per chapter	3821.3
Average number of tokens per section	1008.82

3.2. LARI Dataset

For model evaluation, we used the LARI *dataset*[2] [Junqueira et al. 2026], developed specifically for educational *Question Answering* tasks in Brazilian Portuguese. The dataset was derived directly from the BPLN NLP textbook and consists of context-question-answer triples related to the NLP domain.

Beyond the construction of the triples, LARI also includes a human evaluation stage conducted by the textbook authors, in which each instance received scores ranging from 1 to 5 according to quality and pedagogical adequacy criteria.

Unlike automatically translated *datasets*, LARI was originally created in Portuguese and validated by human specialists in the field. We used exclusively the questions that received the highest score (5), resulting in 238 instances out of a total of 464. This decision was made to reduce potential ambiguities in the evaluation set and to ensure that the benchmark consisted only of instances considered to have the highest pedagogical quality by the dataset creators. Consequently, the evaluation focuses on high-quality reference questions, making the comparison between models more reliable and reducing the influence of potential noise in the benchmark. The *dataset* was employed solely for evaluation purposes and was not used during any stage of model training or adaptation. This separation between adaptation and evaluation data was intentionally designed to prevent data leakage and to assess the models ability to generalize beyond the instructional material seen during training.

3.3. Evaluated Models

We select Qwen3 0.6B, Qwen3 8B, and Sabiá 7B models to investigate how model scale, linguistic specialization, and adaptation strategies influence educational QA tasks in Portuguese. The Qwen3 models were selected because they are open models belonging to the same architectural family [Yang et al. 2025], allowing a more controlled comparison across different parameter scales. While Qwen3 0.6B represents a compact and computationally accessible model, Qwen3 8B offers greater capacity for textual representation and reasoning. The model family also includes multilingual support and capabilities targeted at generative and text understanding tasks.

We also evaluated Sabiá 7B, a commercial model developed for Brazilian Portuguese [Pires et al. 2023b]. Its inclusion makes it possible to analyze the effects of linguistic specialization in tasks involving technical NLP concepts in Portuguese [Assis et al. 2025].

[2]https://huggingface.co/datasets/jjuliar/plndataset

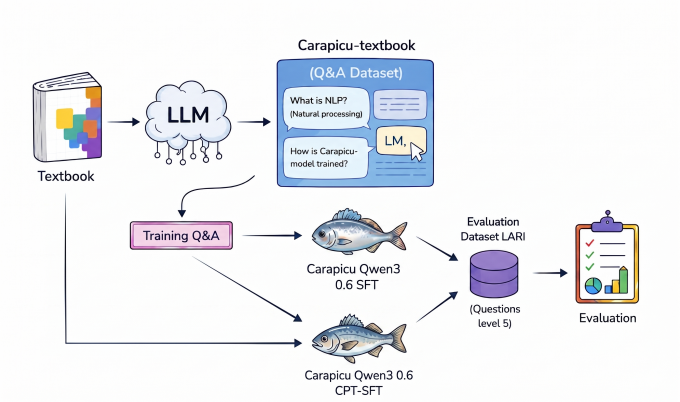


Figure 1. Methodological pipeline adopted in this study.

Beyond the base models, we evaluated the performance of Qwen3 0.6B within a RAG pipeline built on top of the BPLN NLP textbook content, using Qwen3 Embedding 0.6B for question vectorization and Qwen3 Reranker 0.6B to retrieve the most semantically related textbook sections.

4. Methodology: Carapicu Pipeline

Figure 1 presents an overview of the proposed *pipeline*. The BPLN NLP textbook [Caseli and Nunes 2024] was used as the sole knowledge source for the construction of the *carapicu-textbook* dataset, which was later employed for model adaptation through supervised fine-tuning and, in one variant, continued pretraining.

After the adaptation stage, the models were evaluated exclusively using LARI *dataset* questions that received the maximum score (5), resulting in a total of 238 triples. The generated answers were analyzed using BERTScore and an *LLM-as-a-Judge* setup, in which Sabiá-4 evaluated factual correctness, relevance, and completeness on a Likert scale ranging from 1 to 5. The following subsections describe the procedures adopted in each stage of the *pipeline*.

4.1. *Dataset Carapicu-textbook*

The *Carapicu-textbook* dataset was constructed exclusively from the third edition of the book *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português* [Caseli and Nunes 2024], which served as the sole knowledge source. The use of a single educational resource allows us to investigate how language models behave when exposed to structured and domain-specific knowledge in Portuguese. Unlike heterogeneous corpora automatically collected from the Web, the textbook presents stronger conceptual consistency, terminological standardization, and pedagogical organization.

The textual content was initially extracted and organized in JSON format while preserving the original chapter and section structure. Each section was then submitted to the GPT-4o-mini API together with a *prompt*[3] designed to generate 20 question-answer pairs divided into three thematic groups: (i) foundations and concepts; (ii) techniques and models; and (iii) applications and challenges.

[3]https://github.com/FORMAS/carapicu-model

The resulting resource, named *Carapicu-textbook*, consists of a collection of question-answer pairs derived from the textbook content and was used during the model adaptation stage. Table 2 presents general statistics of the *dataset*.

In addition to the automatic generation of question-answer pairs, we conducted a qualitative analysis of the *Carapicu-textbook* dataset through manual inspection of a representative sample. This analysis aimed to verify conceptual consistency, semantic alignment, and adherence to the content of the BPLN NLP textbook. Overall, the generated pairs exhibited thematic coherence and alignment with the concepts presented in their respective sections. We also observed a predominance of short and objective answers, a characteristic that was later reflected in the behavior of the adapted models. Unlike the LARI dataset, whose reference answers are generally more detailed, the Carapicu-textbook emphasizes concise responses. This difference should be taken into account when interpreting the completeness scores reported in the evaluation. Although a full manual validation of the entire dataset was not performed, the sampling-based review allowed us to identify occasional inconsistencies and assess the overall quality of the resource used in the experiments.

Table 2. General statistics of the Carapicu-Textbook dataset

Metric	Value
Total number of QA pairs	1,875
Total number of *dataset* tokens	220,726
Average number of user tokens (question)	18.47
Average number of model tokens (answer)	40.25

4.2. Model Adaptation

The model adaptation stage was designed to investigate how different specialization strategies affect small language models in educational QA tasks in Portuguese. For this purpose, we selected Qwen3 0.6B, an SLM with lower computational cost and inference time, using the *Carapicu-textbook* dataset as the main adaptation resource.

In this work, we considered two strategies: supervised fine-tuning (SFT) and *continued pretraining* (CPT) followed by SFT (CPT-SFT), both applied to Qwen3 0.6B.

In the first strategy, named Carapicu-Qwen3-0.6B-SFT, the model was fine-tuned using the question–answer pairs from *Carapicu-textbook*. Hyperparameter configuration was obtained through Bayesian Optimization and HyperBand (BOHB) [Falkner et al. 2018] using the Weights & Biases (WandB)[4] platform.

In the second strategy, named Carapicu-Qwen3-0.6B-CPT-SFT[5], we conducted a CPT stage using the raw textbook content in order to expose the model to NLP terminology and concepts before supervised fine-tuning. The resulting model was then submitted to the same SFT procedure using the question–answer pairs from *Carapicu-textbook*.

The two strategies allow a comparison between a model directly adapted for QA and another previously exposed to the textual domain of the technical textbook before

[4]`https://wandb.ai`

[5]https://huggingface.co/FORMAS/Carapicu-Qwen3-0.6B-CPT-SFT

supervised fine-tuning. Table 3 presents the final hyperparameters obtained during the optimization process, allowing for the reproduction of the configuration used in both adaptation strategies.

Table 3. Training hyperparameters of the Carapicu models

model	lr	epochs	batch size	eval/loss
Carapicu-Qwen3-0.6B-SFT	3e-5	4	2	1.50
Carapicu-Qwen3-0.6B-CPT-SFT7	3e-5	3	2	1.58

4.3. Retrieval-Augmented Generation

We also developed a *Retrieval-Augmented Generation* (RAG) pipeline to verify whether Qwen3 0.6B would be able to answer adequately when provided with the textbook sections that were semantically closest to the query. We adopted an *Advanced RAG* approach incorporating a *reranker* model to refine the retrieved passages before inference.

The vector database was built using ChromaDB, all sections of the BPLN NLP textbook were indexed and embedded using Qwen3 Embedding 0.6B. We then employed Qwen3 Reranker 0.6B to reorder the retrieved passages: initially, 10 sections were retrieved, and the 3 highest-ranked sections were used as context for answer generation.

During inference, we used a *prompt*[6] instructing the model to answer exclusively based on the provided context. We also set the temperature to 0 and limited the generated responses to a maximum of 512 *tokens*.

4.4. Evaluation

The answers generated by the models were evaluated using BERTScore [Zhang et al. 2019] and a qualitative assessment based on the *LLM-as-a-Judge* paradigm [Lin et al. 2021]. This strategy was adopted to analyze not only textual similarity between generated and reference answers, but also aspects related to conceptual adequacy, relevance, and completeness of the responses produced by the models. Since reference answers are available, traditional automatic metrics can be reliably used for model ranking, as prior work has shown strong alignment between these metrics and human judgments in ranking settings [Thakur et al. 2024].

4.4.1. BERTScore

As the main automatic evaluation metric, we adopted BERTScore [Zhang et al. 2019], an approach based on contextual *embeddings* that measures semantic similarity between texts using representations produced by Transformer models. Unlike metrics based solely on lexical overlap, such as BLEU [Papineni et al. 2002] and ROUGE [Lin 2004], BERTScore can capture semantic similarity even in scenarios involving paraphrases and textual reformulations.

In this work, BERTScore was computed using BERTimbau [Souza et al. 2020] to compare the answers generated by the models against the reference answers from the LARI *dataset*, allowing us to examine the semantic alignment of the produced responses.

[6] https://github.com/FORMAS/carapicu-model

4.4.2. Sabiá-4 as Judge

In addition to automatic metrics, we adopted an *LLM-as-a-Judge* strategy [Zheng et al. 2023], in which a language model acts as an automatic evaluator of the generated responses. Sabiá-4 was the judge model because it was specifically developed for Brazilian Portuguese, offering stronger linguistic and terminological alignment for the evaluation of educational answers in Portuguese [Pires et al. 2023a]. The use of an external model also helps avoid self-evaluation bias [Xu et al. 2024].

Sabiá-4 assigned scores on a Likert scale [Likert 1932] ranging from 1 to 5 for three criteria: (i) factual correctness; (ii) answer relevance with respect to the question; and (iii) completeness. During the qualitative analysis, we observed that the adapted models predominantly generated short and objective answers, reflecting the characteristics of the *Carapicu-textbook* used during training. In contrast, the LARI evaluation dataset generally contains more detailed reference answers. Therefore, the lower completeness scores observed for the adapted models likely reflect differences between the response style learned during adaptation and the style expected by the evaluation benchmark, rather than an actual lack of conceptual knowledge.

5. Results

Table 4 presents the results obtained with BERTScore. In general, the models adapted using the *Carapicu-textbook* dataset achieved the highest semantic similarity scores when compared to the non-specialized base models. The best performance was obtained by Carapicu-Qwen3-0.6B-CPT-SFT, which reached an F1 score of 0.5354, followed by Carapicu-Qwen3-0.6B-SFT with an F1 score of 0.5313. These results indicate that both supervised fine-tuning and the additional continued pretraining stage contributed to bringing the generated answers closer to the reference content contained in the LARI *dataset*.

The results also indicate that smaller models adapted to the educational domain represented by the BPLN NLP textbook can outperform larger non-specialized models. Although Qwen3-8B has substantially higher parametric capacity, its BERTScore performance remained below that of the adapted models, reaching an F1 score of 0.3978. This behavior supports the idea that supervised adaptation using structured instructional knowledge may exert a stronger influence on answer alignment than simply increasing the parameter scale of the model. It is also worth mentioning that the maximum generation length during inference was fixed at 512 *tokens*. Under this restriction, Qwen3-8B frequently exhausted the available budget while producing its reasoning chain (through *think* blocks), leading to truncated outputs and preventing the model from generating a final answer.

We also observed that the RAG pipeline achieved competitive performance in terms of semantic *recall*, obtaining the highest *recall* value among all evaluated models (0.5823). This behavior suggests that retrieving passages from the BPLN NLP textbook favors broader semantic coverage of the generated answers, although this does not always translate into the highest overall alignment measured by F1.

The results obtained by Sabiá-7B indicate that linguistic specialization alone is insufficient to guarantee stronger performance in specialized educational *Question Answering* tasks. Although the model was developed specifically for Brazilian Portuguese, its

BERTScore performance remained below that of the models adapted with the *Carapicu-textbook* dataset. During qualitative analysis, we observed signs of textual degeneration, including repetitive loops, semantically incoherent sentences, and the generation of random characters. It is worth noting that the evaluated version corresponds to the first public release of the Sabiá family available on Hugging Face.

Table 4. Average BERTScore values obtained by the evaluated models

Model	Precision	Recall	F1
Carapicu-Qwen3-0.6B-CPT-SFT	0.5267	0.5526	0.5354
Carapicu-Qwen3-0.6B-SFT	0.5200	0.5513	0.5313
Answer_RAG	0.4529	0.5823	0.5028
Qwen3-0.6B	0.3995	0.5443	0.4575
sabia-7b	0.4284	0.4938	0.4534
Qwen3-8B	0.3453	0.4825	0.3978

Table 5 presents the results of the qualitative evaluation based on the *LLM-as-a-Judge* approach. Unlike BERTScore, which measures semantic alignment with respect to the reference answers, this evaluation examines aspects related to *factual correctness*, *relevance*, and *completeness* of the generated responses.

The results show that the adapted models achieved stronger performance mainly in the relevance dimension. Carapicu-Qwen3-0.6B-CPT-SFT obtained the highest average score for this criterion (3.92), followed by Carapicu-Qwen3-0.6B-SFT (3.85).

In the factual correctness dimension, the retrieval-based *pipeline* achieved the best overall result (3.30), indicating a greater ability to produce factually consistent answers. The adapted models also obtained competitive performance in this criterion, outperforming the base models from the Qwen3 family.

With respect to completeness, we observed more modest results for the adapted models. During qualitative analysis, we found that these models frequently produced shorter and more objective answers, a behavior possibly associated with the characteristics of the *Carapicu-textbook* dataset used during training. For this reason, short answers were not automatically considered inadequate during qualitative evaluation as long as they were conceptually sufficient to answer the proposed question correctly.

We also observed that larger models did not necessarily achieve stronger qualitative performance. Despite its higher parametric capacity, Qwen3-8B obtained the lowest average scores among the evaluated Qwen models across all three analyzed criteria.

Table 5. *LLM-as-a-Judge* results using Sabiá-4. Answers between a Likert scale ranging from 1-strongly disagree to 5-strongly agree.

model	factual correctness	SD	relevance	SD	completeness	SD
Answer_RAG	3.30	±1.56	3.76	±1.62	3.16	±1.46
Qwen3-0.6B	2.78	±1.50	3.34	±1.73	2.67	±1.40
Qwen3-8B	2.13	±1.78	2.08	±1.75	1.85	±1.42
Carapicu-Qwen3-0.6B-SFT	2.97	±1.31	3.85	±1.34	2.70	±1.10
Carapicu-Qwen3-0.6B-CPT-SFT	3.04	±1.28	**3.92**	±1.29	2.79	±1.11
Sabia-7B	1.31	±0.92	1.40	±1.11	1.17	±0.61

6. Discussions

Our results indicate that *fine-tuning* with structured instructional knowledge positively affected the semantic alignment of the generated answers. Both Carapicu models achieved the highest BERTScore F1 values, outperforming the base models and Sabiá-7B, suggesting that direct exposure to the textbook content contributed to bringing the generated answers closer to the LARI reference responses at the semantic level. The results also indicate that small models can benefit considerably from specialization strategies when the domain is well delimited and the training material presents consistent conceptual organization. The comparison between Qwen3-0.6B and the Carapicu models reinforces this interpretation, since all models share the same architectural foundation.

Another aspect observed in the experiments is that increasing parameter scale did not necessarily lead to better performance. Qwen3-8B obtained lower results than the adapted models and even lower scores than its smaller counterpart, Qwen3-0.6B. This behavior suggests that larger models may require a greater number of *tokens* to structure their reasoning chains. In practice, we observed that Qwen3-8B frequently exhausted the entire generation budget during internal processing (*think* blocks), resulting in truncated and incomplete answers. This finding carries critical implications for resource-constrained environments. In local deployment scenarios, where GPU memory, power budget, and latency bounds are strictly limited, larger models not only incur higher computational overhead, but their tendency to produce verbose internal reasoning steps can actively degrade usability.

The results obtained with Sabiá-7B support the interpretation that specialized fine-tuning plays a central role in this scenario. Although the model was natively trained for Brazilian Portuguese, its performance remained below that of the Carapicu variants, indicating that linguistic specialization alone does not guarantee the conceptual alignment required by a technical domain. In addition, qualitative analysis of the Sabiá-7B outputs revealed severe textual degeneration issues. In many interactions, the model produced anomalous outputs characterized by repetitive loops, semantically incoherent sentences, and random character generation.

The *LLM-as-a-Judge* evaluation complements the BERTScore results by revealing qualitative aspects that are not fully captured by traditional metrics. The Carapicu models achieved stronger performance mainly in the relevance dimension, indicating closer alignment between the generated answers and the proposed questions. This behavior is consistent with the adopted training procedure, since *Carapicu-textbook* was constructed directly from question-answer pairs derived from the textbook content.

With respect to completeness, the adapted models achieved lower results than those observed for relevance and factual correctness. One possible explanation lies in the characteristics of the *Carapicu-textbook dataset* used during training. Since this *dataset* was constructed from objective question–answer pairs, most of its answers are relatively short and focused on the core information. Consequently, the adapted models learned to reproduce this pattern during inference, favoring concise responses even when the *LARI dataset* contained more detailed reference answers. Therefore, the lower completeness scores appear to reflect differences between the writing style learned during adaptation and the response format expected by the evaluation benchmark, rather than indicating lower conceptual quality of the generated answers.

It is also important to note that the performance achieved by the RAG strategy in this evaluation may indicate a complementary relationship between retrieval-based methods and parametric adaptation in the studied task, suggesting a viable path for developing higher-quality systems with lower computational costs.

In general, our results indicate that *fine-tuning* with structured instructional knowledge is a promising strategy for adapting small language models to educational QA within the specific NLP domain investigated in this work. Since the experiments were conducted using a single textbook as the knowledge source, further studies are necessary to verify whether these findings generalize to other educational domains and instructional materials.

7. Conclusion

In this work, we investigated how different adaptation strategies influence small language models in educational QA tasks in Portuguese. Our findings indicate that the adapted models achieved stronger semantic alignment and higher answer relevance than the evaluated non-specialized base models. Within the investigated setting, Carapicu-Qwen3-0.6B-SFT and Carapicu-Qwen3-0.6B-CPT-SFT outperformed larger models across different evaluation metrics, suggesting that supervised adaptation with structured instructional knowledge can be an effective strategy for domain-specific educational QA.

The experiments also indicate that the characteristics of the training dataset directly influence the textual behavior of the adapted models, particularly regarding answer objectivity. These findings reinforce the importance of carefully designing instructional datasets for model specialization. More broadly, this work contributes to ongoing discussions on LLM adaptation in Portuguese, educational QA, and the use of structured instructional knowledge for specializing small language models.

Future work should investigate whether the observed behavior remains consistent when other textbooks, knowledge domains, educational resources, and languages are considered. We also plan a human evaluation of the questions in the Carapicu-textbook dataset, aiming to identify the types of questions that language models answer well or struggle with. Based on a taxonomy of question types, we intend to investigate how different categories affect model performance and use these findings to support the practical adoption of this resource as educational material within the BPLN textbook. We also plan to extend the dataset and experiments to the recently released fourth edition of the BPLN textbook.

Limitations

Some limitations of this work should be described. First, the experiments relied exclusively on the BPLN NLP textbook as the knowledge source, restricting the analysis to a specific educational domain within Natural Language Processing. Therefore, the findings should be interpreted within the investigated educational setting and do not necessarily generalize to other textbooks, educational resources, or knowledge domains.

Another limitation is related to the size and characteristics of the *Carapicu-textbook* dataset. Since the resource is composed predominantly of short and objective answers, the adapted models tended to reproduce this pattern during inference, particularly affecting the completeness results.

In addition, the evaluation based on the *LLM-as-a-Judge* paradigm carries limitations inherent to the use of language models as automatic evaluators, potentially reflecting preferences associated with writing style or answer length [Chen et al. 2024]. Furthermore, the judge model employed in this study belongs to the *Sabiá* family, while one of the evaluated models also belongs to the same family. This setup may introduce *self-recognition bias*, a phenomenon in which models tend to favor outputs generated by themselves or by closely related model families [Panickssery et al. 2024]. Although the *Sabiá*-based judge did not explicitly demonstrate systematic preference toward answers generated by the *Sabiá-7B* model, it is possible that some degree of this bias may still have influenced the evaluation results. In future work, we will complement this approach using human evaluation.

Finally, the experiments were conducted only with models from the Qwen3 family and with Sabiá-7B. Future work may extend this analysis by considering other architectures, adaptation strategies, and additional educational datasets in Portuguese.

Ethical Considerations Statement

This work relies exclusively on public academic textual resources for the construction of the *datasets* and the execution of the experiments. The content used for model adaptation was derived from the BPLN NLP textbook, licensed under the Creative Commons BY-NC-ND 3.0 BR license and used in this work with permission from the editors, while the LARI *dataset* was employed solely for evaluation purposes, maintaining an explicit separation between training and testing data. The experiments did not involve the collection of personal or sensitive data. We acknowledge, however, that language models may produce incorrect or biased answers. For this reason, the presented results should be interpreted as experimental analyses within a specific educational setting rather than as a replacement for instructional materials or specialized human evaluation.

Statement on the Use of Generative AI

Generative Artificial Intelligence tools were used as support during the development of this work, including assistance with grammatical revision and translation from Portuguese into English. All methodological decisions, experimental analyses, result interpretations, and final content validation, however, were carried out by the authors. Language models were also employed as part of the proposed experimental methodology, specifically during the evaluation stage based on the *LLM-as-a-Judge* paradigm.

Acknowledgments

The authors acknowledge the support provided by the undergraduate research scholarship from UFBA (2025/2026 PIBIC program), the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 (process no. 88887.970806/2024-00), CNPq (grant 307088/2023-5), FAPERJ (processes SEI-260003/002930/2024, SEI-260003/000614/2023) and the Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação (INCT-TILD-IAR) (grant 408490/2024-1).

References

Asai, A., Wu, Z., Wang, Y., Sil, A., and Hajishirzi, H. (2024). Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. *arXiv preprint arXiv:2310.11511*.

Assis, G., Freitas, C., and Paes, A. (2025). Exploring brazil's llm fauna: Investigating the generative performance of large language models in portuguese. *Journal of the Brazilian Computer Society*, 31(1):939–971.

Caseli, H. M. and Nunes, M. G. V., editors (2024). *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN, 3 edition.

Caseli, H. M. and Nunes, M. G. V., editors (2026). *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN, 4 edition.

Chen, G. H., Chen, S., Liu, Z., Jiang, F., and Wang, B. (2024). Humans or LLMs as the judge? A study on judgement bias. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Costa, Leandro e Souza-Filho, J. B. d. O. e. (2024). Adaptação de llms a novos domínios: um estudo comparativo de estratégias de ajuste fino e rag para tarefas de qa em português. In Claro, Daniela Barreiro e Pagano, A., editor, *Anais do 15º Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, Belém do Pará, Brasil. Associação para Linguística Computacional.

Falkner, S., Klein, A., and Hutter, F. (2018). Bohb: Robust and efficient hyperparameter optimization at scale. In *International conference on machine learning*, pages 1437–1446. PMLR.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., and Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C. C. T., Del Giorno, A., Gopi, S., Javaheripi, M., Kauffmann, P., de Rosa, G., Saarikivi, O., et al. (2023). Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., and Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Junqueira, J. d. R., Freitas, L. A. d., and Corrêa, U. B. (2026). LARI dataset: A native Portuguese question answering dataset from brasileiras em PLN. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors,

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1, pages 1055–1061, Salvador, Brazil. Association for Computational Linguistics.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 140:1–55.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Lin, S., Hilton, J., and Evans, O. (2021). Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. (2024). Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., and Hajishirzi, H. (2023). When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. *arXiv preprint arXiv:2212.10511*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. (2022). Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Panickssery, A., Bowman, S. R., and Feng, S. (2024). Llm evaluators recognize and favor their own generations. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C., editors, *Advances in Neural Information Processing Systems*, volume 37, pages 68772–68802. Curran Associates, Inc.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., and Miller, A. (2019). Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Pires, R., Abonizio, H., Almeida, T., and Nogueira, R. (2023a). Sabiá: Portuguese large language models. In *Anais da XII Brazilian Conference on Intelligent Systems*, pages 226–240, Porto Alegre, RS, Brasil. SBC.

Pires, R., Abonizio, H., Almeida, T. S., and Nogueira, R. (2023b). Sabiá: Portuguese large language models. *arXiv preprint arXiv:2304.07880*.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In Su, J., Duh, K., and Carreras, X., editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Roberts, A., Raffel, C., and Shazeer, N. (2020). How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426. Association for Computational Linguistics.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Thakur, A. S., Choudhary, K., Ramayapally, V. S., Vaidyanathan, S., and Hupkes, D. (2024). Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. *arXiv preprint arXiv:2406.12624*.

Xing, W., Nixon, N., Crossley, S., Denny, P., Lan, A., Stamper, J., and Yu, Z. (2025). The use of large language models in education. *International Journal of Artificial Intelligence in Education*, 35(2):439–443.

Xu, W., Zhu, G., Zhao, X., Pan, L., Li, L., and Wang, W. (2024). Pride and prejudice: Llm amplifies self-bias in self-refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15474–15492.

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., and Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1):90–112.

Yang, A., Li, A., Yang, B., et al. (2025). Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., et al. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Towards Uniform Meaning Representation for Brazilian Portuguese: Building a First UMR-Annotated Dataset

Adriana S. Pagano[1], **Magali Duran**[2], **Federica Gamba**[3], **Daniel Zeman**[3]

[1] Faculdade de Letras – Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte – Brazil

[2]Instituto de Ciências Matemáticas e de Computação
Universidade Federal de São Carlos (UFSCar)
São Carlos – Brazil

[3]Institute of Formal and Applied Linguistics – Charles University
Prague – Czech Republic

apagano@ufmg.br, magali.duran@gmail.com,
{gamba,zeman}@ufal.mff.cuni.cz

Abstract. *This paper presents the first Brazilian Portuguese dataset annotated for Uniform Meaning Representation (UMR). It contains 96 sentences from the Brazilian Portuguese portion of the Parallel Universal Dependencies treebank, parallel to existing UMR annotations in English, Czech, and Italian. The sentences were parsed with PortParser, revised in Arborator-Grew, converted from CoNLL-U into preliminary sentence-level UMR graphs, and manually revised in PENMAN format. The results show that CoNLL-U is a useful starting point, but semantic graph construction requires manual interpretation and language-specific lexical resources. The dataset expands multilingual UMR coverage and supports future Portuguese semantic annotation.*

1. Introduction

As natural language processing moves beyond surface-level linguistic analysis toward deeper forms of meaning representation, semantic annotation has become increasingly important for the development of interpretable, cross-linguistic, and semantically robust language technologies. This need is especially relevant for less represented languages, for which semantically annotated resources remain limited. Brazilian Portuguese is one such case: although it is represented in syntactic resources such as Universal Dependencies [Rademaker et al. 2017, Pardo et al. 2021], there are still few resources that encode sentence-level meaning in a structured graph format.

Uniform Meaning Representation (UMR) offers a framework for building such resources, since it supports cross-linguistic semantic annotation while representing phenomena such as predicate-argument structure, aspect, modality, reference, temporal relations, and reported content [Van Gysel et al. 2021]. Recent work has also emphasized the development of UMR as a broader multilingual infrastructure, including resources and tools for annotation across languages [Bonn et al. 2024].

This paper presents the first UMR-annotated dataset for Brazilian Portuguese. The dataset consists of 96 sentences selected from the Brazilian Portuguese portion of the Parallel Universal Dependencies treebank [Zeman et al. 2017]. These sentences were selected because their counterparts had already been annotated in UMR for English, Czech,

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 285–296
October 19-22, 2026 ©2026 Association for Computational Linguistics

and Italian. As a result, the Brazilian Portuguese annotation can be integrated into an existing multilingual UMR resource rather than standing as an isolated dataset.

The paper makes two main contributions. First, it provides a pilot UMR dataset for Brazilian Portuguese, contributing to the development of semantic resources for Portuguese and to the broader multilingual expansion of UMR. Second, it analyzes the revision process itself, showing which aspects of meaning can be partially derived from CoNLL-U and which require manual semantic interpretation.

The remainder of the paper is organized as follows. Section 2 introduces the UMR framework and related work on multilingual semantic annotation. Section 3 describes the source data and the conversion and revision workflow. Section 4 characterizes the resulting Brazilian Portuguese UMR dataset and analyzes the main types of revision. Section 5 discusses the implications of the dataset for Portuguese semantic annotation and for future UD-to-UMR conversion. Section 6 concludes the paper and outlines next steps.

2. Background and Related Work

Uniform Meaning Representation (UMR) is a graph-based semantic annotation framework designed to support cross-linguistic meaning representation [Van Gysel et al. 2021]. It builds on Abstract Meaning Representation (AMR), originally developed primarily for English [Banarescu et al. 2013], while extending its scope to typologically diverse and under-resourced languages. At the sentence level, UMR represents predicate-argument structure, semantic roles, entity reference, modality, aspect, quantification, and named entities. At the document level, it also provides mechanisms for representing coreference, temporal relations, and modal dependencies [Bonn et al. 2024]. These features make UMR suitable for multilingual semantic annotation, since it provides a shared representational layer while allowing language-specific grammatical and lexical distinctions to be captured.

A central challenge for UMR is the effort demanded by annotation. As with other forms of deep semantic annotation, producing UMR graphs requires linguistic expertise and careful manual work. For this reason, recent work has explored strategies for bootstrapping UMR annotation from already available linguistic resources. One approach is to convert AMR corpora into UMR [Bonn et al. 2023]. This relies on the fact that UMR inherits many representational principles from AMR. However, relying only on AMR-to-UMR conversion limits the development of UMR resources to languages for which AMR annotations already exist. This is problematic for multilingual expansion, especially in the case of less represented languages.

Other initiatives have explored conversion from different types of linguistic resources. [Buchholz et al. 2024] propose bootstrapping UMR from interlinear glossed text, generating preliminary UMR structures for Arapaho from language documentation resources. Another line of work converts semantically richer dependency resources, such as the Prague Dependency Treebank, into UMR [Lopatková et al. 2025]. The tectogrammatical layer of the Prague Dependency Treebank includes deep syntactic-semantic information, which provides richer input for conversion than basic dependency trees. However, this type of resource is not available for most languages.

The UD-to-UMR approach proposed by [Gamba et al. 2025] is relevant to our

work because it aims at a language-independent conversion pipeline. Universal Dependencies (UD) provides broad multilingual coverage and a relatively consistent morphosyntactic annotation scheme across languages [de Marneffe et al. 2021]. Although UD is not a semantic representation, its dependency structures encode information about predicates, dependents, grammatical relations, and morphological features that can be partially mapped to UMR structures. In the UD-to-UMR converter, concept nodes are generally derived from lemmas, while participant roles are assigned through linguistically informed rules. For example, UD relations such as `nsubj`, `csubj`, and `obl:agent` may be mapped to UMR `:actor`, while `obj`, `nsubj:pass`, and `csubj:pass` may be mapped to `:undergoer`. Morphological information can also contribute to semantic interpretation; for instance, dative obliques may be interpreted as recipients.

The resulting graph is not intended to be a complete UMR representation, but rather a structure that annotators can revise and complete. This is important because many UMR distinctions cannot be reliably inferred from syntax alone. Named entity types, abstract predicates, reentrancies, implicit arguments, fine-grained aspectual distinctions, and discourse-level relations require lexical-semantic knowledge or manual interpretation. [Gamba et al. 2025] therefore treat conversion as a way to facilitate annotation rather than as a fully automatic semantic analysis.

3. Methodology

This section describes the construction of the Brazilian Portuguese UMR dataset. The goal was to produce a manually revised semantic dataset while assessing which aspects of UMR annotation can be supported by syntactic input.

The source data consist of 96 sentences selected from the Brazilian Portuguese portion of the Parallel Universal Dependencies treebank (PUD) [Zeman et al. 2017]. These sentences were selected because their corresponding English, Czech, and Italian sentences had already been annotated in UMR. This design makes the Portuguese dataset compatible with an existing multilingual UMR resource and allows the annotation process to benefit from parallel semantic analyses.

The Brazilian Portuguese sentences were first automatically parsed with PortParser ([Lopes and Pardo 2024] and subsequently revised in Arborator-Grew ([Guibon et al. 2020]. This process produced CoNLL-U files containing tokenization, lemmas, part-of-speech tags, morphological features, dependency relations, and syntactic heads. These revised CoNLL-U files served as the morphosyntactic input for the initial conversion into UMR.

The first stage consisted in automatically converting the CoNLL-U sentences into preliminary UMR graphs. The conversion followed the general UD-to-UMR strategy proposed by [Gamba et al. 2025]. In this approach, the converter creates preliminary UMR concept nodes from the lemmas in the CoNLL-U file. Thus, inflected word forms are initially represented through their base forms. Dependency relations and morphological features provide cues for building the initial graph structure -dependency relations provide cues for participant structure, and morphological features contribute to the assignment of semantic attributes where possible.

In the initial graphs, verbal and nominal predicates were represented as concept nodes whenever they could be identified from the syntactic structure. Core syn-

tactic dependents were mapped to preliminary semantic roles. For example, subjects and agents were generally mapped to actor-like roles, while objects and passive subjects were mapped to undergoer-like roles. Other dependents, such as obliques and modifiers, were often assigned broader placeholder relations when a more specific UMR relation could not be reliably inferred from syntax alone.

The conversion also introduced placeholders for categories that require manual interpretation. For instance, named entities could be detected from proper-name structures, but their semantic type, such as person, organization, or place, often required manual correction. Similarly, aspectual values could not always be inferred from UD morphology and were therefore revised manually. This is consistent with the purpose of UD-to-UMR conversion: the converted graph provides a useful starting point, but several semantic distinctions require annotator intervention.

After conversion, the preliminary UMR graphs were manually revised through direct text editing in PENMAN format. This allowed the annotator to inspect and modify the Penman-style graph structure, correct concept labels, normalize semantic roles, resolve placeholder values, and add missing semantic information.

Manual revision focused on five main areas. First, predicate-argument structures were revised: subjects, objects, passive constructions, oblique arguments, control structures, and reported clauses were checked in relation to the predicate involved. We also examined eventive nouns and verbal periphrases, since these constructions often require decisions that go beyond syntactic dependency structure. Eventive nouns were evaluated to determine whether they introduced event concepts in the UMR graph, while verbal periphrases were analyzed to decide which verbs should be represented as semantic predicate nodes and which should instead be captured through aspectual annotation.

Second, named entities were typed and normalized, including the use of `:name` structures and appropriate entity labels. Third, aspect and modality values were revised, since these categories often depend on lexical and constructional interpretation. Fourth, temporal, spatial, and discourse-related relations were checked, especially when the initial graph contained generic placeholder relations. Fifth, parallel UMR annotations in Italian were consulted when useful, while preserving language-specific semantic decisions for Brazilian Portuguese.

The revision process also included the treatment of implicit or underspecified participants. In some sentences, Portuguese allows arguments to remain unexpressed when they are recoverable from context or from the construction. In these cases, the graph was revised to include the semantic participant when required by the UMR analysis.

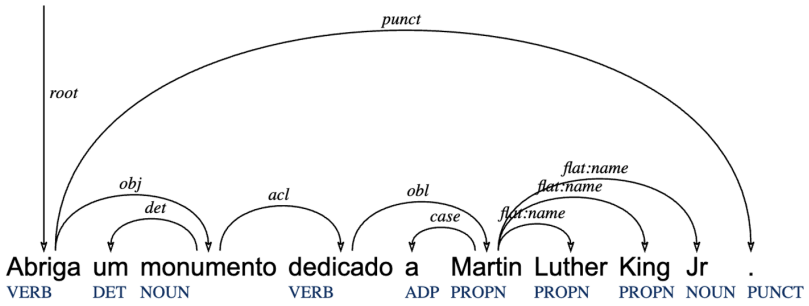
A central part of the revision process involved the normalization of predicate-specific argument roles. While the initial conversion produced general participant roles such as `:actor`, `:undergoer`, and related semantic labels, the manually revised graphs were normalized into frame-based `:ARG` roles whenever possible. For this purpose, we used Verbo Brasil [Duran et al. 2013, Duran and Aluísio 2015] as a lexical-semantic resource for Brazilian Portuguese.

For each verbal predicate, we checked whether a corresponding Verbo Brasil frame was available. When a frame existed, its argument structure was used to guide the assignment of UMR roles. When no Verbo Brasil frame was available, or when the avail-

able frame did not adequately cover the sentence, the annotation was completed through manual analysis. In such cases, we considered the predicate meaning, the syntactic structure, and the parallel UMR annotations in Italian. This procedure ensured that the final annotation remained semantically motivated even when lexical resources were incomplete.

Figure 1 illustrates our workflow with the sentence: "Abriga um monumento dedicado a Martin Luther King Jr." The CoNLL-U analysis identifies *Abriga* as the root predicate, *monumento* as its object, and *dedicado a Martin Luther King Jr.* as an adjectival clausal modifier of *monumento*. The automatic conversion yields a preliminary semantic graph, using broad roles such as `:actor`, `:undergoer`, and `:OBLIQUE`. In the manually revised graph, these relations are normalized into predicate-specific roles: *monumento* is analyzed as `:ARG1` of `abrigar`, and, following the Verbo Brasil roleset for *dedicar*, *monumento* is also `:ARG1` of `dedicar`, while *Martin Luther King Jr.* is represented as `:ARG2`. The revision also replaces placeholder values for named entity type, aspect, and modal strength with `person`, `State`, and `FullAff`.

1. CoNLL-U input



2. Automatically converted UMR graph

```
(s20a / abrigar
  :actor (s20f / FILL
    :refer-number singular)
  :undergoer (s20m / monumento
    :refer-number singular
    :undergoer-of (s20d / dedicar
      :OBLIQUE (s20n / NE-TYPE
        :name (s20n2 / name
          :op1 "Martin"
          :op2 "Luther"
          :op3 "King"
          :op4 "Jr"))
      :modal-strength MS-affirmative
      :aspect ASP))
  :modal-strength full-affirmative
  :aspect ASP)
```

3. Manually revised UMR graph

```
(s20a / abrigar
  :ARG1 (s20m / monumento
    :refer-number singular
    :ARG1-of (s20d / dedicar
      :ARG2 (s20p / person
        :name (s20n / name
          :op1 "Martin"
          :op2 "Luther"
          :op3 "King"
          :op4 "Jr"))
      :modal-strength FullAff
      :aspect State))
  :modal-strength FullAff
  :aspect State)
```

Figure 1. Annotation workflow from CoNLL-U input to automatically converted and manually revised UMR graphs.

To analyze the contribution of manual revision, we compared the initial converted graphs with the final revised graphs. The comparison was qualitative and focused on the main types of changes introduced during revision.

The comparison focused on recurring revision patterns, including: replacement of

general participant roles with predicate-specific :ARG roles; correction of named entity types; resolution of placeholder relations such as generic obliques; addition or removal of participants; normalization of aspect and modality values; restructuring of reported speech and embedded clauses; and treatment of constructions where the syntax-semantics mapping was non-isomorphic. These categories were used to characterize which aspects of the final UMR graphs could be supported by CoNLL-U and which required manual semantic interpretation.

4. Results

This section characterizes the final manually revised Brazilian Portuguese UMR dataset. It reports the main annotation layers, including predicate-specific argument roles, aspect, modality, reported speech, special UMR predicates, and named entities, and briefly comments on the revision patterns that shaped the final graphs.

The first version of the dataset was produced through automatic conversion from CoNLL-U into preliminary UMR graphs. This conversion provided a useful starting point for annotation, since it identified candidate predicates, introduced initial concept nodes, and mapped some syntactic dependents to general semantic roles. In particular, syntactic relations such as subjects, objects, passive subjects, and obliques were converted into broad participant or modifier relations.

The automatically converted graphs thus contained general participant labels, such as :actor, :undergoer, and :theme, as well as placeholder values for categories that require semantic interpretation. These included generic aspect labels, modal-strength values, named entity types, and broad oblique relations.

Table 1 provides a general overview of the final Brazilian Portuguese sentence-level UMR dataset. The corpus contains 96 sentence-level UMR graphs, corresponding to 1,970 word tokens when punctuation is excluded and 2,216 tokens when punctuation is included. The table also summarizes key semantic annotation layers in the dataset, including 559 predicate-specific ARG relations, 11 reported-speech relations encoded with :quot, and 56 occurrences of special UMR predicates. In addition, the corpus contains 123 named entities, 71 of which are linked to external identifiers through :wiki values.

Dataset feature	Count
Sentence-level UMR graphs	96
Word tokens, excluding punctuation	1,970
Word tokens, including punctuation	2,216
Predicate-specific ARG relations	559
Reported-speech relations encoded with :quot	11
Special UMR predicates	56
Named entities	123
Named entities linked through :wiki values	71

Table 1. Overview of the Brazilian Portuguese sentence-level UMR dataset.

To characterize the predicate-argument layer of the corpus, Table 2 reports the distribution of predicate-specific ARG roles after manual revision. Counts combine equiva-

lent direct and inverse realizations of the same role, since both encode the same participant function in the graph.

ARG role	Count	% of ARG relations
ARG0	129	23.1
ARG1	289	51.7
ARG2	112	20.0
ARG3	24	4.3
ARG4	5	0.9
Total	**559**	**100.0**

Table 2. Distribution of predicate-specific argument roles in the Brazilian Portuguese sentence-level UMR dataset. Counts combine direct and inverse realizations of the same ARG role.

The distribution shows that ARG1 is the most frequent role, accounting for more than half of all predicate-specific argument relations. This reflects the high number of undergoer, theme, or affected-participant roles in the corpus. ARG0 and ARG2 are also well represented, while ARG3 and ARG4 are comparatively rare and occur mainly in predicates with more specialized argument structures.

Aspect and modality are central components of UMR annotation and were therefore checked during manual revision. Table 3 presents the distribution of aspectual values in the manually revised dataset. The most frequent values are State and Performance, with 115 occurrences each, followed by Activity with 48 and Process with 18. The remaining aspectual values occur less frequently.

Aspect value	Frequency
State	115
Performance	115
Activity	48
Process	18
Endeavor	2
Generic	2
Habitual	1

Table 3. Aspect values in the manually revised UMR dataset.

The distribution indicates a balance between stative and eventive predications. The high frequency of State reflects descriptions, classifications, and property attributions, while Performance and Activity account for dynamic situations. These distinctions are not directly available in the original CoNLL-U annotation and therefore required manual revision.

Table 4 shows the distribution of modal-strength values. FullAff is by far the most frequent value, with 256 occurrences. Negative, neutral, and partial modal values are less frequent, but their presence shows that the annotation captures variation in polarity and modal commitment.

The concept inventory contains 625 unique concept labels. Table 5 lists the most frequent concepts in the manually revised dataset. The most frequent concept is name,

Modal strength	Frequency
FullAff	256
FullNeg	17
NeutAff	16
PrtAff	7
PrtNeg	3
NeutNeg	2

Table 4. Modal-strength values in the manually revised UMR dataset.

with 124 occurrences, followed by person, with 81 occurrences. This reflects the high number of named entities and named individuals in the PUD sentences.

Concept	Frequency
name	124
person	81
and	43
date-entity	28
identity-91	20
country	17
have-mod-91	10
thing	10
human-settlement	10
have-role-91	9
after	7
outro	7
event	7
grande	6
muito	6
esse	6
have-rel-role-92	6
include-91	5
ethnic-group	5
company	5

Table 5. Twenty most frequent concepts in the manually revised UMR dataset.

The frequent occurrence of abstract predicates such as identity-91, have-mod-91, have-role-91, have-rel-role-92, and include-91 shows that the revised annotation introduces abstract semantic structures required by UMR.

The main effect of manual revision was the transformation of a general UD-derived graph backbone into a predicate-specific semantic representation. General participant labels introduced during conversion were replaced with :ARG roles whenever possible, guided by Verbo Brasil and by manual analysis of predicate meaning. This is reflected in the high frequency of :ARG0, :ARG1, :ARG2, and :ARG3 in the final dataset.

Manual revision also resolved placeholder values introduced during conversion. Generic aspect labels were replaced by specific aspectual values such as State,

Performance, Activity, and Process. Generic modal labels were replaced by values such as FullAff, FullNeg, NeutAff, and PrtAff. Named entities were typed, :name structures were checked, and :wiki links were added when appropriate.

A further effect of revision was the replacement of broad syntactic-semantic relations with more specific UMR relations. Generic oblique relations were revised as :temporal, :place, :manner, :purpose, :reason, :condition, :source, or :goal, depending on the meaning of the construction. Reported content was also manually revised through :quot structures where required.

Overall, these results show that CoNLL-U provides a useful starting point for UMR annotation, but that the final Brazilian Portuguese dataset required manual semantic analysis and language-specific lexical resources. The final graphs are therefore not merely converted syntactic structures, but semantically enriched UMR representations.

5. Discussion

The Brazilian Portuguese UMR epresents a first step toward the development of deeper semantic graph resources for Portuguese. The manually revised graphs encode predicate-argument structure, reference, aspect, modal strength, named entities, temporal and spatial relations, reported content, and abstract predicates, confirming the usefulness of UMR for representing semantic phenomena that go beyond the morphosyntactic information available in CoNLL-U.

The results show that automatic UD-to-UMR conversion provides a useful annotation starting point, but the final graphs required substantial revision because the mapping from syntax to semantics is often non-isomorphic: syntactic subjects, objects, and obliques do not always correspond directly to predicate-specific semantic roles.

A central revision step was the normalization of predicate-argument structure. General roles such as :actor, :undergoer, and :theme were replaced with predicate-specific :ARG roles whenever possible, supported by Verbo Brasil. Other layers, including aspect, modality, named entity typing, :wiki links, abstract predicates, and relations such as :temporal, :place, :manner, and :purpose, also required manual interpretation beyond dependency labels.

Overall, the workflow combines CoNLL-U input, automatic conversion, manual revision, and language-specific lexical-semantic resources. It offers a practical model for developing UMR datasets for languages with syntactic resources but limited semantic graph annotation.

6. Conclusion

This paper presented the first UMR-annotated dataset for Brazilian Portuguese: 96 sentence-level graphs from the Brazilian Portuguese portion of the Parallel Universal Dependencies treebank, manually revised after automatic UD-to-UMR conversion. Because the sentences are parallel to existing UMR annotations in English, Czech, and Italian, the dataset contributes to a broader multilingual UMR resource.

Future work will expand the dataset, refine the use of Brazilian Portuguese lexical resources, compare the Portuguese graphs more systematically with their English, Czech,

and Italian counterparts, improve conversion rules for Portuguese-specific constructions, and release the dataset as part of a multilingual UMR infrastructure.

6.1. Limitations

This study is limited by the small size of the dataset, which contains 96 sentence-level graphs. The annotations were carried out collaboratively by the authors, rather than through a blind procedure, due to the incipient stage of UMR annotation for Brazilian Portuguese and the lack of language-specific guidelines. The dataset is therefore best understood as a pilot resource and a first step toward larger-scale Brazilian Portuguese UMR annotation.

7. Acknowledgments

The authors express their gratitude to the anonymous reviewers for their constructive feedback. Adriana Pagano's research is funded by FAPEMIG and the National Council for Scientific and Technological Development (CNPq, grants 404722/2024-5; 313103/2021-6). Magali Sanches Duran is a member of the Artificial Intelligence Center of the University of São Paulo, funded by the São Paulo Research Foundation (FAPESP grant no. 2019/07665-4) and IBM. She is also supported by the Ministry of Science, Technology and Innovations, with resources from Law No. 8,248 of October 23, 1991, within the scope of PPI-Softex, coordinated by Softex and published as ICT Residency 13, DOU 01245.010222/2022-44. This work received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology) and is conducted under the auspices of the National Institute of Science and Technology in Responsible Artificial Intelligence for Computational Linguistics, Information Treatment, and Dissemination (INCT-TILDIAR), funded by CNPq (grant 408490/2024-1).

8. Ethics disclosure

This work uses sentences from the publicly available Parallel Universal Dependencies treebank. The dataset does not contain private, sensitive, or newly collected personal data. The Brazilian Portuguese UMR graphs were produced for research purposes through automatic conversion and manual revision. The resulting resource is intended to support linguistic research and the development of multilingual semantic annotation, and will be released with appropriate attribution to the source data and in accordance with the relevant dataset licenses.

9. Generative AI use

The authors declare that no generative AI or AI-assisted technologies were used in the preparation of this manuscript.

References

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N. (2013). Abstract Meaning Representation for sembanking. In Pareja-Lora, A., Liakata, M., and Dipper, S., editors, *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Bonn, J., Buchholz, M. J., Chun, J., Cowell, A., Croft, W., Denk, L., Ge, S., Hajic, J., Lai, K., Martin, J. H., et al. (2024). Building a broad infrastructure for uniform meaning representations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2537–2547.

Bonn, J., Myers, S., E. L. Van Gysel, J., Denk, L., Vigus, M., Zhao, J., Cowell, A., Croft, W., Hajič, J., H. Martin, J., Palmer, A., Palmer, M., Pustejovsky, J., Urešová, Z., Vallejos, R., and Xue, N. (2023). Mapping AMR to UMR: Resources for adapting existing corpora for cross-lingual compatibility. In Dakota, D., Evang, K., Kübler, S., and Levin, L., editors, *Proceedings of the 21st International Workshop on Treebanks and Linguistic Theories (TLT, GURT/SyntaxFest 2023)*, pages 74–95, Washington, D.C. Association for Computational Linguistics.

Buchholz, M. J., Bonn, J., Post, C. B., Cowell, A., and Palmer, A. (2024). Bootstrapping UMR annotations for Arapaho from language documentation resources. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2447–2457, Torino, Italia. ELRA and ICCL.

de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Duran, M. and Aluísio, S. (2015). Automatic generation of a lexical resource to support semantic role labeling in Portuguese. In Palmer, M., Boleda, G., and Rosso, P., editors, *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 216–221, Denver, Colorado, USA. Association for Computational Linguistics.

Duran, M. S., Martins, J. P., and Aluísio, S. M. (2013). Um repositório de verbos para a anotação de papéis semânticos disponível na web (a verb repository for semantic role labeling bavailable in the web) [in Portuguese]. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*.

Gamba, F., Palmer, A., and Zeman, D. (2025). Bootstrapping UMRs from Universal Dependencies for scalable multilingual annotation. In Peng, S. and Rehbein, I., editors, *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)*, pages 126–136, Vienna, Austria. Association for Computational Linguistics.

Guibon, G., Courtin, M., Gerdes, K., and Guillaume, B. (2020). When collaborative treebank curation meets graph grammars. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5291–5300.

Lopatková, M., Hledíková, H., Štěpánek, J., and Zeman, D. (2025). From the prague dependency treebank to the uniform meaning representation: Gold-standard czech umr data and partial automatic conversion.

Lopes, L. and Pardo, T. (2024). Towards portparser - a highly accurate parsing system for Brazilian Portuguese following the Universal Dependencies framework. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Processing*

of Portuguese - Vol. 1, pages 401–410, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Pardo, T. A. S., Duran, M. S., Lopes, L., Di Felippo, A., Roman, N. T., and Nunes, M. d. G. V. (2021). Porttinari: a large multi-genre treebank for brazilian portuguese. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, Porto Alegre, RS, Brasil. SBC.

Rademaker, A., Chalub, F., Real, L., Freitas, C., Bick, E., and de Paiva, V. (2017). Universal dependencies for portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 197–206.

Van Gysel, J. E., Vigus, M., Chun, J., Lai, K., Moeller, S., Yao, J., O'Gorman, T., Cowell, A., Croft, W., Huang, C.-R., et al. (2021). Designing a uniform meaning representation for natural language processing. *KI-Künstliche Intelligenz*, 35(3):343–360.

Zeman, D., Popel, M., Straka, M., Hajič, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., Tyers, F., Badmaeva, E., Gokirmak, M., Nedoluzhko, A., Cinková, S., Hajič jr., J., Hlaváčová, J., Kettnerová, V., Urešová, Z., Kanerva, J., Ojala, S., Missilä, A., Manning, C. D., Schuster, S., Reddy, S., Taji, D., Habash, N., Leung, H., de Marneffe, M.-C., Sanguinetti, M., Simi, M., Kanayama, H., de Paiva, V., Droganova, K., Martínez Alonso, H., Çöltekin, Ç., Sulubacak, U., Uszkoreit, H., Macketanz, V., Burchardt, A., Harris, K., Marheinecke, K., Rehm, G., Kayadelen, T., Attia, M., Elkahky, A., Yu, Z., Pitler, E., Lertpradit, S., Mandl, M., Kirchner, J., Alcalde, H. F., Strnadová, J., Banerjee, E., Manurung, R., Stella, A., Shimada, A., Kwak, S., Mendonça, G., Lando, T., Nitisaroj, R., and Li, J. (2017). CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In Hajič, J. and Zeman, D., editors, *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

Novo CorPop Saúde: Data Collection for Simplified Health Communication in a Brazilian Patient Information Leaflet *Corpus*

Bianca Pasqualini[1,2], **Maria J. B. Finatto**[1,2], **Eduarda Bertotto**[1]
Paula Salem Carpio[2], **Roges Horacio Grandi**[3], **Leandro Krug Wives**[3]

[1] Fundação de Amparo à Pesquisa Estado do Rio Grande do Sul (FAPERGS)
Porto Alegre – RS – Brazil

[2] PPG-LETRAS – Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre – RS – Brazil

[3] PPGIE – CINTED – Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre – RS – Brazil
{bianca.pasqualini, eduarda.bertotto, mfinatto, roges.grandi,
leandro.wives, paula.carpio}@ufrgs.br

Abstract. *This paper presents exploratory results from the "Novo CorPop Saúde" project, evaluating whether validation by expert linguists improves readability levels compared to AI-generated simplifications. We analyzed original, AI-generated, and human-validated versions of thirty Patient Information leaflets. Using the Flesch Reading Ease Index, two adapted Dale-Chall metrics, and lexicometric analysis, we assessed automatic simplification against human strategies. Results indicate that while AI improves readability scores primarily through text compression, expert linguists prioritize elaborative strategies, maintaining key terminology while providing didactic explanations to fill semantic gaps. This suggests that human intervention plays a key role in balancing terminological precision with communicative accessibility, a nuance that purely reductive AI strategies might overlook. This study provided evidence on the use of the complex and simplified lexical repertoire of medicine leaflets for the "Novo CorPop Saúde".*

1. Introduction

The Novo CorPop Saúde (NCS)[1] is a research and innovation project in Lexicon and Terminology Studies based on Textual and Terminological Accessibility (TTA) and Natural Language Processing (NLP). Its objective is to expand the existing CorPop [Pasqualini 2018][2] by incorporating texts from Health Sciences and Pharmacology to provide resources for simplifying health information to match the country's average literacy

[1]Project conducted at UFRGS. Granted by FAPERGS, Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável (INCT TILD-IAR) and supported by healthcare professionals.

[2]Available at https://www.ufrgs.br/textecc/porlexbras/corpop/index.php.

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 297–304
October 19-22, 2026 ©2026 Association for Computational Linguistics

level.[3] The original CorPop is a reference for written popular Brazilian Portuguese, based on the population's literacy level, comprising diverse genres such as newspapers, literature, and informational brochures, and is used to assess general linguistic accessibility.

The NCS research complies with the Brazilian legal framework, specifically Law 15.263 on plain language[4] and Law 14.874[5], which establish clear communication as a legal obligation and linguistic accessibility as a universal design principle. By affirming the citizen's right to understand official communications, this legal landscape positions accessible health information as a basic right intended to protect patient autonomy and citizenship from technical complexity, especially considering the country's low literacy levels. Importantly, 54.1% of Brazilians struggle to comprehend Patient Information leaflets (PILs), which causes 25% of users to stop their medication because of confusion, despite the fact that 98.4% of the population considers PILs necessary and 59.6% report reading them regularly [Dal Pizzol et al. 2019].

The NCS will include diverse written health materials from official sources: PILs ("bulas de medicamentos"), official educational materials (health booklets, guidelines, and texts directed at the general public produced by Anvisa), informed consent forms ("termos de consentimento livre e esclarecido"), and AI-generated content (a curated sample of health-related texts for laypeople produced by generative AI, such as ParkinsonBot [Campos 2025]).

The project is structured into three main phases: Phase 1 (evaluation and validation of AI simplifications), Phase 2 (healthcare professionals' reviews for scientific and clinical accuracy) and Phase 3 (reading tests with patients to assess real-world impact). In this initial phase (Phase 1), the work is focused on a *corpus* of PILs of Summaries of Product Information for the most consumed drugs nationwide. We analyzed the sample in [Scalercio et al. 2026], which includes all hypertension drugs from the 2025 Programa Farmácia Popular do Brasil[6], as detailed in the next sections.

The analytical approach of this study is grounded in research bridging Textual and Terminological Accessibility (TTA), Health Literacy (HL), and Natural Language Processing (NLP). [Finatto and Tcacenco 2021] define text simplification as intralingual translation, showing that accessibility requires elaborative strategies, such as explanatory paraphrasing, rather than simple reductive summarization. [Finatto and Paraguassu 2022] apply this to the Brazilian context, where high rates of functional illiteracy require the mediation of medical terminology through familiar vocabulary to support patient autonomy. To evaluate these transformations, [Menezes et al. 2023] used NLP for Automatic Readability Assessment (ARA), noting that traditional formulas fail to capture the didactic expansions of human curation. Recently, [Wives and Finatto 2025] and [Scalercio et al. 2025] explored Large Language Models (LLMs) for simplification and conceptual modeling, showing that Generative AI can generate accessible versions, but

[3]According to INAF, 58% of the Brazilian population has rudimentary to elementary literacy levels, creating a significant demand for accessible materials. Available at: https://alfabetismofuncional.org.br/alfabetismo-no-brasil/.

[4]Law 15.263, November 14th 2025.

[5]Law 14.874, May 28th 2024, regulates clinical research involving human subjects and establishes the national system of ethics in research with human beings in Brazil

[6]https://www.gov.br/saude/pt-br/composicao/sectics/farmacia-popular

expert validation is essential for accuracy. It's worth mentioning [Basu et al. 2023] as well, who dealt specifically with controllable simplification of medical texts.

2. Methodology and analysis

The methodology is based on the comparative analysis of 30 hypertension PILs. We established three parallel *corpora* along with a reference *corpus*. The first *corpus*, C1 (original PILs), consists of the original texts provided by the manufacturers. The second, S1 (AI simplification), consists of versions automatically simplified by AI. The third, S2 (human validation), consists of versions reviewed by linguists. Additionally, the CorPop [Pasqualini 2018] was used as the baseline for accessible Brazilian Portuguese vocabulary (Table 1).

Table 1. Description and lexical characterization of the parallel *corpora*.

***Corpus*/set**	**Tokens**	**Types**	**TTR**
CorPop (reference)	688,788	32,311	0.0469
C1 (original PILs)	77,306	3,116	0.0403
S1 (AI simplification)	41,991	2,279	0.0543
S2 (human validation)	48,532	2,513	0.0518

The dataset S1 used ChatGPT (version 5.2) to process the original PILs, guided by a specific set of prompt instructions to: (1) replace technical terminology with simpler synonyms, keeping the original term in parentheses only when unavoidable; (2) convert passive voice constructions into the active voice; (3) split lengthy sentences into shorter, more direct ones; (4) omit irrelevant or excessive details; (5) resolve anaphoras and ambiguous pronouns; and (6) strictly preserve all numerical data, such as dosages, frequencies, and routes of administration[7]. The automatic simplifications (S1) were then refined by three linguists specialized in plain-language techniques to determine validated versions (S2) [Scalercio et al. 2026]. The human validation (S2) was done by linguists who had access to both C1 and S1.

The Flesch Reading Ease Index was calculated using the Vidya Text Tool[8], based on the formula adapted for Brazilian Portuguese [Martins et al. 1996]. In this scale, higher scores indicate lower textual complexity, with values above 75 (easy) being desirable for readers with low literacy. As shown in Table 2, the original PILs (C1) present a very difficult baseline (22.4). While AI simplification (S1) improves this score to 40.9, the human-validated versions (S2) achieve the highest overall readability score (42.8). Despite the improvement, both simplified versions remain within the "difficult" range (equivalent to "university level" of complexity; see Table 2).

To complement the analysis, we used two variations of the Dale-Chall formula [Chall and Dale 1995] adapted for Brazilian Portuguese. The first, Vidya-Dale-Chall (Vidya-DC), uses the technological validation from the Vidya Text Tool. The second, CorPop-Dale-Chall (CorPoP-DC), uses the 3,000 most frequent words from CorPop as its reference for easy words (excluding grammatical words). The results show that S2

[7]This repository contains the code and the resources used: https://github.com/scalercio/med-simple-docs; [Scalercio et al. 2026].

[8]Available at http://vidyatext.nuvem.ufrgs.br/.

Table 2. Flesch Reading Ease Index metrics across the parallel *corpora*.

Corpus/set	Flesch	Interpretation	Vidya-DC	CorPop-DC	Interpretation
C1 (orig. PILs)	22.4	Very difficult	10.08	10.61	Post-graduate level (very difficult)
S1 (AI simp.)	40.9	Difficult	9.63	9.93	University level (difficult)
S2 (hum. valid.)	42.8	Difficult	9.34	9.71	University level (low)

consistently achieves the most accessible scores in both metrics (Table 2). This assessment suggests that human elaborative strategies tend to improve lexical accessibility, as indicated by the consistency across these metrics.

Table 3. Consolidated lexical analysis: frequency, ratio and difference across parallel *corpora*.

Word	S1 (%)	S2 (%)	CorPop (%)	Ratio (S2/CP)	Diff. (S2-S1)
médico	1.329%	1.263%	0.029%	43.9	-0.066%
pressão*	0.867%	0.938%	0.006%	161.4	+0.071%
você	0.914%	0.870%	0.420%	2.1	-0.045%
dose	0.953%	0.870%	0.001%	1197.8	-0.083%
medicamento	0.779%	0.826%	0.002%	569.1	+0.048%
coração*	0.562%	0.777%	0.052%	14.8	+0.215%
sangue*	0.572%	0.680%	0.011%	60.0	+0.108%
usar	0.564%	0.525%	0.014%	38.5	-0.039%
dia	0.641%	0.556%	0.172%	3.2	-0.084%
rins	0.367%	0.451%	0.001%	444.0	+0.085%
tratamento	0.398%	0.398%	0.020%	19.8	0.000%
problemas	0.181%	0.383%	0.049%	7.9	+0.202%
aumentar	0.441%	0.394%	0.007%	54.2	-0.047%
potássio	0.393%	0.354%	0.000%	—	-0.039%
remédios	0.324%	0.330%	0.004%	90.8	+0.006%

*Usage predominantly metaphorical in CorPop.

To evaluate the lexical choices made during the simplification process, we compared the relative frequencies of key terms (excluding grammatical words and multi-word expressions, making sure that the analysis captures the primary lexical shifts) across S1, S2, and CorPop. Table 3 presents a consolidated analysis of the fifteen most frequent content words, which were selected by calculating the relative frequency of unigrams in the S2 dataset, along with relevant terminological units (such as “rins” and “potássio”). By cross-referencing these high-frequency tokens across the parallel *corpora*, we can observe how human curation adjusts word distribution to match medical terminology with the accessible vocabulary patterns of the CorPop baseline.

The data indicate that the human elaborative strategy (S2) increases the frequency of concrete, familiar terms (such as “coração”, “sangue”, “problemas”, and “remédios”) compared to the AI output (S1). Notably, the word “problemas” is used more than twice as frequently in S2 as in S1, a shift of +0.202% that underscores the human preference for high-coverage lexical choices over more specialized words. These terms also show a stronger baseline presence in CorPop, confirming their accessibility to lay readers. In contrast, the linguists tended to reduce the frequency of certain technical or prescriptive terms, like “dose”, “usar”, and “aumentar”, which were more prominent in the AI output.

The qualitative impact of the elaborative strategy is best observed through direct

comparisons of specific sentences. Table 4 provides eight representative examples that highlight the differences between reductive automatic methods and human validation.

Table 4. Microstructural analysis: sentence transformation examples organized by NCS layers.

#	C1 (original PILs)	S1 (AI simplification)	S2 (human validation)
	LEXICAL LAYER		
1	O atenolol é indicado para o controle da hipertensão arterial.	O atenolol serve para controlar a pressão alta.	O atenolol serve para controlar a pressão alta.
2	Este fármaco é excretado primordialmente por via renal.	O remédio sai do corpo pela urina.	O remédio sai do corpo principalmente através da urina (pelos rins).
3	A eficácia deste medicamento depende da posologia correta.	O remédio só funciona se tomar direito.	Para o remédio fazer efeito, você deve tomar as doses nos horários certos.
	DISCURSIVE LAYER		
4	O fármaco atua bloqueando os receptores beta-adrenérgicos.	O remédio age no coração diminuindo a pressão.	O remédio ajuda o coração a bater mais devagar e com menos força.
5	Não tome duas doses ao mesmo tempo.	Não tome duas doses juntas.	Nunca tome duas doses ao mesmo tempo para compensar a que esqueceu.
	PRAGMATIC LAYER		
6	A pressão arterial deve ser aferida regularmente.	Meça a pressão sempre.	Por isso, você deve medir sua pressão regularmente.
7	Podem ocorrer tonturas, fadiga e cefaleia.	Pode causar tontura, cansaço e dor de cabeça.	Você pode sentir tontura, cansaço e dor de cabeça.
8	Pacientes com disfunção renal requerem ajuste posológico.	Quem tem problema nos rins precisa de doses menores.	Pessoas com problemas nos rins podem precisar de doses diferentes.

While AI (S1) focuses on lexical substitution, the human strategy (S2) incorporates discursive and pragmatic layers to help provide clearer comprehension. For instance, in example 5, the human strategy provides meaning to a semantically hollow AI construction ("pressão que cai") by describing the actual physical sensation ("tontura ou vista escura"). In example 6, the addition of the justification "para compensar a que esqueceu" and the emphatic "nunca" provides a more complete functional conceptualization for the reader, helping to avoid confusion. Similarly, the shift from impersonal third-person structures (e.g., "pacientes" and "podem ocorrer") to direct engagement ("você pode sentir") in examples 8 and 9 reflects the strategy of establishing communicative proximity. This approach supports the reader's understanding by personalizing the health information and making the content more accessible.

The strategies shown in these layers of intralingual translation [Finatto and Tcacenco 2021] represent the core elements to be incorporated into the NCS repertoire:

1. **Lexical layer**: technical terms changed into accessible equivalents (such as "pressão alta" instead of "hipertensão"). This also includes specific words and terms that are infrequent in CorPop, such as "medicamento".
2. **Discursive layer (explanatory paraphrases)**: didactic explanations that serve as a reference standard for explaining complex mechanisms, such as how the medicine acts on the heart.
3. **Pragmatic layer**: direct engagement strategies, including the use of "você" and the active voice, which bring the text closer to the reader, as validated by the linguists

3. Discussion

Human validation (S2) yields higher readability than S1 (Table 2) by prioritizing terminological precision through elaboration. The modest gain in the Flesch index (42.8 versus 40.9) is a direct consequence of the strategy to expand the text with didactic explanations. Because these formulas are sensitive to sentence length, adding context through

explanatory paraphrasing limits improvements in the score. Still, S2 achieved better results in both Flesch and Dale-Chall metrics despite its larger token volume (Table 1). This suggests that choosing accessible terms compensates for the longer sentence length, reinforcing the importance of all three layers in the NCS repertoire.

Beyond global scores, the analysis of frequency ratios and lexical ranking (Table 3) is key to determining the terms to be included in the NCS repertoire. A qualitative screening in CorPop shows that, while biological terms like "coração", "sangue", and "pressão alta" are highly frequent, they often appear in metaphorical contexts. However, this high frequency in common language indicates **reinforced familiarity**: because these terms are already deeply integrated into the reader's vocabulary, they serve as ideal anchors for intralingual translation [Finatto and Tcacenco 2021]. By prioritizing familiar biological terms over specialized terminology (such as "hipertensão"), this selection process takes advantage of existing linguistic habits to reduce cognitive load and improve accessibility. This strategy is particularly relevant because reading difficulties tend to intensify under stress, such as in self-medication contexts, where linguistic complexity may contribute to 25% of users stopping their medication prematurely [Dal Pizzol et al. 2019].

In contrast, the analysis also identifies essential terms with low frequency in CorPop, such as "medicamento" and "dose" (Table 3). These units exhibit extreme frequency ratios: "medicamento" is 569.1 times more frequent in the PILs (S2) than in the general reference *corpus*, while "dose" is nearly 1,200 times more frequent (ratio 1197.8). In context, these are **terminological units** that carry technical concepts and are candidates for the NCS repertoire. Their consistent presence in S2 (Table 3) highlights that even units from common language require conceptual disambiguation to prevent errors, particularly when they are potentially polysemic [Finatto 2022]. Within the scope of the simplification process as a kind of intralingual translation, these terms form essential parts of the NCS vocabulary, and the discursive layer intentionally offers explanations to bridge semantic gaps for the reader.

4. Final remarks and future work

The analysis of the hypertension PILs sample suggests that human elaborative strategies, organized into lexical, discursive, and pragmatic layers, helped to resolve semantic and structural gaps often found in the AI simplifications. These strategies will be incorporated into the NCS repertoire as the basis of future validation efforts in the next phases of the project. Expanding the NCS will also require a standardized validation protocol for consistency across linguists. This will guide the adaptation of specialized terminology based on the CorPop reference while preserving clinical accuracy through layered explanations. We also intend to investigate multi-word expressions (MWEs) to better capture complex medical concepts in simplified health communications.

As mentioned in previous sections, this study only reports on the initial stages of the project (Phase 1 - PILs). Its conclusions are limited to describing and analyzing the validation strategies applied by linguists to review the automatically generated simplifications. This serves to support the evaluation of the potential use, in the new *corpus*, of AI-generated simplifications and reviewed simplifications. The next stage involves a final technical review by healthcare professionals (Phase 2) so that, in the future *corpus*, the

simplified language alternatives also have scientific quality and conceptual accuracy.

5. Acknowledgments

This research was supported by CNPq (National Council for Scientific and Technological Development), grant PQ 307088/2023-5, National Institutes of Science and Technology – CNPq-TILD-IAR (grant 408490/2024-1), PROBIC FAPERGS-UFRGS, and FAPERGS Ed. 06/2025 - PROGRAMA PESQUISADOR GAÚCHO – PqG (grant 25/2551-0002702). We thank the Programa de Pós-Graduação em Letras/UFRGS for the support. We also thank Silvana de Jesus and Felipe L. Duarte Pacheco, who helped review some simplifications, and Arthur Scalercio, who provided the data files used in this research.

6. Limitations

This exploratory study is limited to a sample of 30 hypertension Patient Information Leaflets. The observed validation strategies and lexical shifts may not fully generalize to other medical domains. As previously reiterated, the human-validated simplifications (S2) have not yet been empirically tested with lay readers to assess real-world comprehension improvement.

7. Ethics statement

This study analyzed public Patient Information Leaflets from Anvisa. No human subjects, personal data, or private medical records were involved. The research adheres to ethical standards for linguistic studies, aiming to promote health literacy and social inclusion.

Disclaimer

In compliance with Portaria CNPq 2.664/2026, Gemini CLI version 0.38.2 was used for processing lexical counts, structural organization, and stylistic refinement. ChatGPT (version 5.2) was used to generate the initial automatic simplifications (S1) as part of the research methodology. GPT was used to check text orthography, style, and grammar.

References

Basu, C., Vasu, R., Yasunaga, M., and Yang, Q. (2023). Med-EASi: Finely Annotated Dataset and Models for Controllable Simplification of Medical Texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12579–12587.

Campos, M. N. (2025). Parkinsonbot: chatbot para auxílio e acessibilidade sobre a doença de parkinson. Trabalho de conclusão de curso (graduação em ciência da computação), Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre. Accessed: 25 Sep. 2025.

Chall, J. S. and Dale, E. (1995). *Readability revisited: the new Dale-Chall readability formula*. Brookline Books.

Dal Pizzol, T. d. S., Moraes, C. G., Arrais, P. S. D., Bertoldi, A. D., Ramos, L. R., Farias, M. R., Oliveira, M. A., Tavares, N. U. L., Luiza, V. L., and Mengue, S. S. (2019). Medicine package inserts from the users' perspective: are they read and understood? *Revista Brasileira de Epidemiologia*, 22:e190009.

Finatto, M. J. B. and Paraguassu, L. B. (2022). Comunicação acessível em temas de saúde pública: textos e terminologias sobre epidemias, do passado ao presente. In Moreira, G. L., Costa, L. A. C., and Alves, I. M., editors, *Pesquisas em Lexicologia, Lexicografia e Terminologia*, volume 1, pages 189–323. Pontes Editores, Campinas.

Finatto, M. J. B. and Tcacenco, L. M. (2021). Tradução intralinguística, estratégias de equivalência e acessibilidade textual e terminológica. *TradTerm*, 37(1):30–63.

Martins, T. B. F., Ghiraldelo, C. M., Nunes, M. G. V., and Oliveira Jr., O. N. (1996). Readability formulas applied to textbooks in brazilian portuguese. *Notas do ICMSC-USP, São Carlos*.

Menezes, L. C., Paes, A., and Finatto, M. J. B. (2023). Abordagem baseada em aumento de dados para avaliação automática de leiturabilidade. *Domínios de Lingu@gem*, 17:e1721. DOI: 10.14393/DLv17a2023-21.

Pasqualini, B. (2018). Corpop - um corpus do português popular escrito. Thesis, Programa de Pós-Graduação em Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre. Accessed: 25 May 2026.

Scalercio, A., Bertotto, E., Jesus, S., Finatto, M. J. B., and Paes, A. (2026). Annotation guidelines and challenges for automatic simplification of portuguese drug leaflets. In *Proceedings of Propor/Lang4Health*.

Scalercio, A., de Souza, E., Finatto, M. J., and Paes, A. (2025). Evaluating LLMs for Portuguese Sentence Simplification with Linguistic Insights. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24452–24477. Association for Computational Linguistics.

Wives, L. K. and Finatto, M. J. B. (2025). Usando LLMs para simplificar e representar documentos médicos. In *Anais Estendidos do XL Simpósio Brasileiro de Bancos de Dados (SBBD 2025)*, pages 415–425, Porto Alegre, RS, Brasil. SBC.

Legal Nugget Extraction for Granular Retrieval over Long Jurisprudential Texts

Lucas Pereira[1], **Erick Brito**[2], **Roberto Lotufo**[3], **Jayr Pereira**[2]

[1]Tribunal de Contas da União (TCU)

[2]Universidade Federal do Cariri (UFCA)

[3]NeuralMind.ai

jayr.pereira@ufca.edu.br

Abstract. *Legal retrieval over jurisprudential collections is challenging because court decisions are long, heterogeneous documents whose relevant legal thesis may occupy only a small portion of the text. This paper asks whether legal nuggets, defined as short and self-contained legal theses extracted from source documents, can improve dense retrieval over Brazilian legal collections. We propose a pipeline that extracts nuggets from each document, indexes them with embeddings, retrieves nugget-level evidence, and aggregates the retrieved nuggets back to document-level rankings. We evaluate this approach on four Portuguese legal retrieval benchmarks from the JUA ecosystem, reporting NDCG@10, MAP@10, and MRR@10. Nugget retrieval substantially improves the two jurisprudential datasets: on JUA-Juris, NDCG@10 increases from 0.10265 to 0.20461, and on JurisTCU from 0.20898 to 0.32696. However, it underperforms full-document retrieval on NormasTCU and BR-TaxQA, and an embedding-model ablation shows that strong domain-adapted retrievers can remain better in the full-document setting. The results demonstrate that legal nuggets can be useful for jurisprudence search, especially when queries are formulated as legal theses, but they may not transfer equally well to other legal retrieval scenarios.*

1. Introduction

Searching jurisprudence is a central task in legal practice because lawyers, judges, auditors, and researchers often need to find prior decisions that support, distinguish, or contradict a specific legal thesis. Unlike many web documents, jurisprudential texts are long, formal, and argumentative: a single ruling may contain procedural history, factual background, references to parties, multiple legal issues, and one or more generalizable holdings. As a result, the document that should be retrieved is often relevant because of a small thesis-bearing fragment rather than because of its full textual content. Dense retrieval has become a central paradigm for modern information retrieval and question answering [Karpukhin et al. 2020, Thakur et al. 2021], but embedding a complete jurisprudential document into a single vector may dilute the signal of the legal thesis that matters for the query.

The idea of retrieving smaller information units has appeared in several contexts. Passage retrieval and retrieval-augmented generation commonly decompose long

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 305–316
October 19-22, 2026 ©2026 Association for Computational Linguistics

documents into retrievable chunks [Lewis et al. 2020], while late-interaction models such as ColBERT preserve more granular matching signals inside the retrieval model [Khattab and Zaharia 2020]. In question answering and evidence-centered retrieval, short atomic pieces of information can also serve as evidence units that are easier to rank, inspect, and combine than full documents. Recent RAG work has made this explicit through *information nuggets*: minimal facts, claims, or question-answer units used to evaluate answer completeness, preserve provenance, guide generation, and construct retrieval benchmarks [Łajewska and Balog 2025, Dietz et al. 2026, Thakur et al. 2025]. These approaches share a common intuition: retrieval may improve when the indexed or evaluated unit is closer to the unit of information requested by the query.

This paper explores legal nuggets as one possible granular representation for jurisprudential search. We define a legal nugget as a short, self-contained statement that captures an atomic legal argument or thesis consolidated by a document. Instead of indexing the full document text directly, our method extracts nuggets from each document, embeds the nuggets, searches over the nugget index, and then maps the retrieved nuggets back to their original documents. The motivation is that many jurisprudential queries resemble legal theses rather than complete case narratives. If the indexed representation is also thesis-like, retrieval may become more robust to noise. At the same time, granular indexing introduces risks: extraction errors may remove important context, and ranking several nuggets per document requires aggregation back to the document level. This makes nugget retrieval a promising but empirical hypothesis, not an automatic replacement for full-document retrieval.

In this context, our research question is: *Can legal nuggets improve dense retrieval over long jurisprudential texts in Brazilian legal collections?* For answering this question, we compare a full-document dense retrieval baseline with a nugget-based index in which each retrieved nugget is aggregated back to its source document.

We evaluate this idea on four datasets from the JUA legal retrieval ecosystem [Pereira et al. 2026a]: JUA-Juris, JurisTCU [Fernandes et al. 2026], NormasTCU, and BR-TaxQA [Júnior et al. 2026]. The experiments compare the nugget-based index against a full-document dense retrieval baseline using the same embedding model. The results indicate a clear split. Nugget retrieval substantially improves results on jurisprudential datasets, especially when queries are ementas or legal theses. In contrast, it underperforms full-document retrieval on NormasTCU and BR-TaxQA in terms of NDCG@10 and MAP@10, although MRR@10 improves in NormasTCU. This mixed behavior is consistent with prior observations that legal NLP benefits from domain-aware representations, but that such representations must be matched to the task and document genre [Chalkidis et al. 2020].

The contribution of this work is threefold: (i) we propose a simple pipeline for legal nugget extraction and nugget-level dense retrieval; (ii) we provide an empirical comparison across four Portuguese legal retrieval tasks; and (iii) we analyze when granular legal representations help, arguing that the benefit depends on the alignment between prompt, document genre, and query type.

2. Related Work

Dense retrieval represents queries and documents as vectors and ranks documents by vector similarity. Dual-encoder models such as DPR [Karpukhin et al. 2020] showed that dense retrieval can outperform sparse term matching in open-domain question answering, while BEIR [Thakur et al. 2021] highlighted the importance of evaluating retrieval methods across heterogeneous tasks. Neural retrieval architectures have also explored more granular forms of matching. ColBERT [Khattab and Zaharia 2020], for example, performs late interaction over token-level embeddings rather than compressing a whole document into a single vector.

Granularity is also central to retrieval-augmented generation, where systems typically retrieve passages or chunks instead of entire documents [Lewis et al. 2020]. Chunking reduces input length and can improve local semantic matching, but generic fixed-size chunks may not correspond to meaningful legal units. Legal texts are highly structured and argument-driven: the most useful unit may be a legal thesis, rule application, or precedent statement rather than an arbitrary window of tokens [Ma et al. 2024].

Information nuggets offer a complementary granular representation for retrieval and evaluation. Recent work uses nuggets as minimal facts, claims, or question-answer units for grounded generation and RAG evaluation. GINGER extracts, clusters, ranks, and summarizes information nuggets to improve grounded response generation [Łajewska and Balog 2025], while Q&A nuggets have been used to preserve citation provenance and guide selection in RAG systems [Dietz et al. 2026]. Nuggets have also been used to construct realistic retrieval benchmarks with nugget-level support judgments [Thakur et al. 2025] and to compare LLM-based relevance judgment methods [Arabzadeh and Clarke 2025]. In the legal domain, Farzi et al. use factual nuggets to help legal professionals compare and improve AI-generated summaries of legal depositions [Farzi et al. 2026]. These works motivate nuggets as interpretable evidence units, but they do not directly evaluate nugget indexing for jurisprudential document retrieval.

Prior work on legal NLP has emphasized that legal language and legal tasks require domain-aware representations [Chalkidis et al. 2020]. In legal retrieval, this suggests that the document representation should reflect the type of legal information being sought. Our work follows this intuition by using large language models to extract legal theses from long jurisprudential documents before embedding them for retrieval.

3. Legal Nugget Retrieval

Figure 1 summarizes the proposed pipeline. The method has two phases. In the offline phase, a corpus of jurisprudential documents is converted into a nugget index: documents are passed to a legal nugget generator, the generated nuggets are linked to their source document identifiers, and an embedding model maps each nugget to a vector stored in a vector index. In the online phase, a user query is embedded with the same embedding model, the vector index retrieves the most similar nuggets, and a document aggregator converts nugget-level evidence into a document-level ranking.

3.1. Nugget Extraction

Nugget extraction is the step that turns a long legal document into a set of smaller semantic units. In this work, a legal nugget is not merely a passage or a fixed-size chunk. It is an

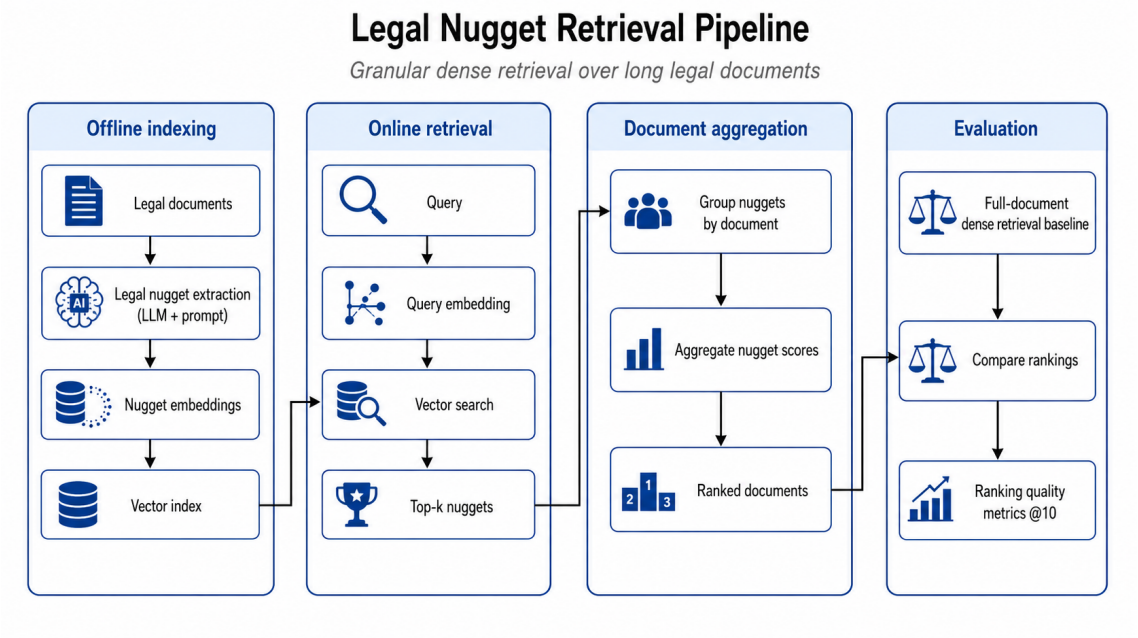


Figure 1. Legal nugget retrieval pipeline. The offline phase converts legal documents into an indexed nugget collection, while the online phase retrieves nuggets for a query and aggregates them into a document ranking.

atomic statement of legal content that can be searched independently from the document that contains it. This follows recent nugget-based work in retrieval and RAG, where compact information units are used to represent factual evidence, guide selection, and preserve provenance [Łajewska and Balog 2025, Dietz et al. 2026]. For jurisprudence, we define nuggets as reusable legal theses or arguments: compact statements that connect a legal rule or principle, a relevant factual pattern, and the conclusion consolidated by the decision.

This definition is motivated by the structure of jurisprudential documents. Judicial and administrative rulings often contain procedural history, identification of parties, factual background, citations, several legal issues, and one or more holdings. A query, however, may target only one of these legal holdings. Indexing the full document as a single unit can therefore dilute the matching signal, especially when the relevant thesis is surrounded by long context that is useful for legal interpretation but less useful for first-stage retrieval.

The extraction step aims to preserve the part of the document that is most likely to be reused in retrieval: the legal proposition itself. The prompt asks the generator to avoid procedural details, isolated citations, party names, document identifiers, purely factual descriptions, and generic dispositive statements. The resulting nugget should be self-contained enough to be matched against a query without requiring the full document, while still maintaining its provenance through the identifier of the source document. Figure 2 shows a simplified version of the jurisprudence-oriented prompt used to operationalize this definition.

You receive the text of a jurisprudential summary or ruling. Extract all atomic legal arguments consolidated by the decision. Each nugget must express one reusable legal thesis connecting a legal rule or principle, a factual pattern, and a conclusion.

Return only legal arguments. Ignore procedural history, party names, document identifiers, isolated citations, purely factual descriptions, and generic dispositive statements. Each nugget must be self-contained and understandable without the full document.

Return a valid JSON object containing a list of nugget texts. Each text is later linked to the source document identifier by the indexing pipeline.

Figure 2. Simplified version of the jurisprudential legal nugget extraction prompt used for JUA-Juris and JurisTCU. Other datasets may use task-specific prompts.

3.2. Indexing and Search

Once extracted, nuggets become the retrievable units of the collection, in a process similar to nugget-based retrieval and generation pipelines that use atomic information units to guide selection and preserve provenance [Dietz et al. 2026, Łajewska and Balog 2025]. This changes the retrieval problem from matching a query against complete legal documents to matching it against candidate legal propositions. The intuition is that a query about a legal thesis may be closer to one specific nugget than to the full decision that contains it. In this setting, the document remains the unit of evaluation and citation, but the nugget acts as the first-stage evidence used to find that document.

Indexing nuggets therefore separates two roles that are conflated in standard document retrieval. The nugget represents the semantic content to be matched, while the source document represents the legal authority to be returned. This distinction is important for long jurisprudential texts: a decision may contain several legal propositions, and only one of them may correspond to the information need expressed by the query. A nugget index allows these propositions to compete individually before the system decides which documents to return.

The final step maps nugget-level evidence back to document-level rankings. For each query, the system first retrieves the top N nuggets from the vector index, where $N = 10{,}000$ in our experiments. Each retrieved nugget has a similarity score and a source document identifier. Documents are then deduplicated by aggregating all retrieved nuggets that point to the same source document. In the main experiments, we use a first-evidence aggregation strategy: because retrieved nuggets are processed in decreasing similarity order, the score assigned to a document is the score of its highest-ranked retrieved nugget. Equivalently, for a query q, document d, retrieved nuggets n_i, scores s_i, and source mapping $doc(n_i)$, the document score is $S(q, d) = \max\{s_i \mid doc(n_i) = d\}$ over the retrieved nugget set. The final ranking is obtained by sorting documents by $S(q, d)$, so each source document appears at most once.

This aggregation choice treats a single highly relevant legal proposition as sufficient evidence for retrieving its source document. We also implemented alternative strategies, such as averaging the top k nugget scores from the same document, which favor documents supported by multiple strong pieces of evidence. These choices are part of the

retrieval model, not merely implementation details, because they determine how granular evidence is translated into the document ranking expected by legal search benchmarks.

4. Experimental Setup

4.1. Datasets

We evaluate on four Portuguese legal retrieval datasets from the JUA benchmark ecosystem [Pereira et al. 2026a]:

- **JUA-Juris**: a jurisprudential retrieval task introduced as part of JUA. The corpus is derived from curated TCU jurisprudence excerpts, and each query is based on an *enunciado*, i.e., a concise statement of the legal thesis associated with a ruling. Relevance is binary and centered on the exact query–document pair, making this dataset especially close to thesis-to-jurisprudence retrieval.
- **JurisTCU**: a jurisprudential corpus from the Brazilian Federal Court of Accounts with query relevance judgments [Fernandes et al. 2026]. The original collection contains curated jurisprudence records with metadata and textual fields such as `ENUNCIADO` and `EXCERTO`, which correspond to the summary of a ruling and the supporting decision excerpt. Its queries include real keyword searches and synthetic queries derived from ruling summaries, with relevance judgments that were manually verified.
- **NormasTCU**: a retrieval task over normative documents from the same legal-administrative domain. Unlike jurisprudential collections, these documents are normative acts that establish general rules and are typically organized into provisions such as articles and paragraphs. In JUA, documents are represented by combining the subject summary field with the full text of the norm, and relevance is graded.
- **BR-TaxQA**: a tax-law question answering retrieval benchmark with references for Brazilian personal income tax law [Júnior et al. 2026]. It is based on FAQ-style questions and answers from the Brazilian Federal Revenue Service. The retrieval task is question-driven: systems must retrieve the answer or reference material associated with a taxpayer question, with graded relevance reflecting primary and secondary linked answers.

The main target of this study is jurisprudential retrieval. Accordingly, JUA-Juris and JurisTCU use the jurisprudence-oriented nugget prompt described in Section 3, which focuses on legal theses consolidated by rulings. NormasTCU and BR-TaxQA are included as out-of-genre checks: they allow us to test whether the same granular retrieval idea transfers beyond jurisprudence to normative and question-driven legal search. For these two datasets, we use slightly different prompts, adapted to extract atomic legal facts from normative texts and tax question–answer material rather than jurisprudential holdings.

4.2. Models and Metrics

Nuggets are generated with `gpt-5.4-nano`. This model is used only for the extraction step: it reads each source document and produces the nugget texts that are later indexed. Retrieval is performed by embedding the generated nuggets and the queries with dense

embedding models; the generator itself is not used at search time. The implementation, run configurations, and prompts are available in a public repository.[1]

The main comparison uses `text-embedding-3-small` for both the full-document baseline and the nugget index. Baseline scores are taken from the public JUA leaderboard[2], which reports standardized results for the benchmark described by Pereira et al. [Pereira et al. 2026a]. The nugget runs use the same embedding model, but the indexed units are extracted nuggets rather than full documents. We use this proprietary embedding model as the main experimental setting because it provides a public, standardized JUÁ leaderboard baseline and allows us to isolate the effect of changing the retrieval unit from complete documents to legal nuggets under the same encoder. The open-weight models are therefore treated as an ablation: they test whether the observed behavior persists when the embedding model changes, rather than serving as the primary point of comparison. We index vectors with FAISS exact search over L2-normalized embeddings. In the ablation, we additionally evaluate Qwen3 Embedding 4B and 8B [Zhang et al. 2025], as well as JUA-adapted 4B retrievers [Pereira et al. 2026b]. These choices instantiate the generic pipeline described in Figure 1; the pipeline itself does not depend on these specific models.

We report standard retrieval metrics at cutoff 10: NDCG@10, a position-aware metric that discounts lower-ranked hits and supports graded relevance [Järvelin and Kekäläinen 2002]; MAP@10, which summarizes precision at ranks where relevant documents appear [Manning et al. 2008]; and MRR@10, which emphasizes the rank of the first relevant hit [Voorhees and Tice 2000].

5. Results and Discussion

Table 1 reports the main comparison across the four datasets.

Table 1. Main retrieval results comparing full-document indexing with legal nugget indexing across four Brazilian legal datasets. Both approaches use the same embedding model in each dataset.

Dataset	Approach	NDCG@10	MAP@10	MRR@10
JUA-Juris	Full document	0.10265	0.07713	0.07732
JUA-Juris	Legal nuggets	**0.20461**	**0.15756**	**0.15756**
JurisTCU	Full document	0.20898	0.08728	0.44645
JurisTCU	Legal nuggets	**0.32696**	**0.17368**	**0.58299**
NormasTCU	Full document	**0.26250**	**0.15815**	0.42702
NormasTCU	Legal nuggets	0.24928	0.11562	**0.49169**
BR-TaxQA	Full document	**0.74126**	**0.65848**	**0.76473**
BR-TaxQA	Legal nuggets	0.65566	0.56618	0.66025

The jurisprudential datasets show the strongest gains. On JUA-Juris, nugget retrieval nearly doubles NDCG@10, from 0.10265 to 0.20461, and also approximately

[1] `https://github.com/ufca-llms/nuggets-search`

[2] `https://huggingface.co/spaces/ufca-llms/jua-leaderboard`

doubles MAP@10 and MRR@10. On JurisTCU, NDCG@10 increases from 0.20898 to 0.32696, while MAP@10 almost doubles from 0.08728 to 0.17368.

The results are different for NormasTCU and BR-TaxQA. In NormasTCU, nuggets slightly underperform full-document retrieval in NDCG@10 and MAP@10, but improve MRR@10. This suggests that the nugget index may move the first relevant result earlier while weakening the overall top-ranked ordering. In BR-TaxQA, the full-document baseline remains stronger across all reported metrics.

5.1. Interpretation

The three metrics highlight different retrieval behaviors. NDCG@10 is position-aware and sensitive to graded relevance across the top ranks; MAP@10 summarizes how consistently relevant documents appear throughout the top-10 ranking; MRR@10 focuses only on the first relevant hit. Reading the metrics together is therefore important: an improvement in MRR@10 without improvement in NDCG@10 or MAP@10 means that the system often finds one relevant document earlier, but does not necessarily produce a better ranked list overall.

The main pattern is that legal nuggets help most when the query and the extracted representation describe the same kind of legal unit. In JUA-Juris, queries are ementas, i.e., summaries of legal theses. The nugget prompt was also designed to extract legal theses from jurisprudence. This alignment turns retrieval into a comparison between thesis-like queries and thesis-like document representations, reducing the noise introduced by long jurisprudential documents.

The JurisTCU result reinforces this interpretation. Even though the nugget index was partly transferred from JUA-Juris and completed for missing documents, the representation remains jurisprudence-oriented. It improves NDCG@10, MAP@10, and MRR@10 over full-document embeddings, indicating that the extracted theses capture useful retrieval evidence.

By contrast, NormasTCU and BR-TaxQA are not purely jurisprudential thesis retrieval tasks. NormasTCU is especially informative because nugget retrieval improves MRR@10 but lowers NDCG@10 and MAP@10. This indicates an early-hit effect: nuggets can help surface at least one relevant normative document sooner, but the remaining top-ranked results are less consistently useful than those produced by full-document retrieval. One possible explanation is that normative relevance often depends on broader regulatory context, definitions, exceptions, or the relationship among provisions. Extracting isolated nuggets may preserve some directly matchable statements while weakening the contextual signals needed to rank several relevant norms well. BR-TaxQA shows an even clearer case where nugget extraction is misaligned with the task: tax questions and answers often depend on complete explanatory context, so full-document answer representations remain stronger across NDCG@10, MAP@10, and MRR@10.

These findings point to a more precise claim: nugget retrieval should not be treated as a generic replacement for full-document retrieval. Instead, it is a task-specific representation strategy. It is particularly promising when the information need is granular, legal-argumentative, and aligned with the prompted extraction schema.

5.2. Embedding Model Ablation

We also run an ablation with stronger open-weight embedding models to examine whether nugget retrieval remains useful when the embedding model changes. Table 2 compares full-document retrieval from the JUA leaderboard with nugget retrieval using the same embedding model. We consider two general Qwen3 embedding models [Zhang et al. 2025] and two JUA-adapted 4B models [Pereira et al. 2026b].

Table 2. Embedding model ablation on jurisprudential datasets at cutoff 10.

Dataset	Model	Approach	NDCG	MAP	MRR
JUA-Juris	jua-4B-mixed	Full	**0.29044**	**0.23056**	**0.23026**
JUA-Juris	jua-4B-mixed	Nuggets	0.24561	0.19034	0.19034
JUA-Juris	jua-4B-legal-only	Full	**0.29371**	**0.23336**	**0.23313**
JUA-Juris	jua-4B-legal-only	Nuggets	0.25398	0.19645	0.19645
JUA-Juris	Qwen3-4B	Full	0.19892	0.15181	0.15188
JUA-Juris	Qwen3-4B	Nuggets	**0.22606**	**0.17354**	**0.17354**
JUA-Juris	Qwen3-8B	Full	0.21719	0.16993	0.17002
JUA-Juris	Qwen3-8B	Nuggets	**0.23499**	**0.18070**	**0.18070**
JurisTCU	jua-4B-mixed	Full	**0.36337**	0.16956	**0.64069**
JurisTCU	jua-4B-mixed	Nuggets	0.33752	**0.17977**	0.58679
JurisTCU	jua-4B-legal-only	Full	**0.37520**	**0.17935**	**0.64986**
JurisTCU	jua-4B-legal-only	Nuggets	0.32963	0.17422	0.57892
JurisTCU	Qwen3-4B	Full	0.31094	0.13752	**0.58815**
JurisTCU	Qwen3-4B	Nuggets	**0.32423**	**0.16550**	0.55645
JurisTCU	Qwen3-8B	Full	**0.33175**	0.14897	0.61106
JurisTCU	Qwen3-8B	Nuggets	0.32852	**0.17115**	**0.62319**

This ablation qualifies the main result. Nugget retrieval improves the general Qwen3 embedding models on JUA-Juris, suggesting that granular legal thesis representations can compensate for the lack of task-specific adaptation. However, it does not outperform the JUA-adapted 4B models in the full-document setting. On JurisTCU, nuggets often improve MAP@10 for the general models, but this improvement does not consistently translate into higher NDCG@10 or MRR@10. Thus, the benefit of nugget retrieval depends not only on the document genre and query type, but also on whether the embedding model has already learned a strong document-level representation for the benchmark.

5.3. Limitations

This study has some limitations. First, nugget generation can be costly, depending on the size of the search collection and on the quality and price of the generation model. Future work should evaluate smaller or open-source models for nugget extraction, especially in settings where the corpus is large or frequently updated. Second, the document-level ranking is obtained with simple aggregation strategies over retrieved nuggets. More sophisticated scoring methods could combine nugget evidence, full-document scores, and reranking models. Third, the evaluation is based on retrieval metrics from existing benchmarks. These metrics are appropriate for comparing ranking systems, but they do not directly measure how useful the retrieved legal evidence is for a practitioner. Future work

should therefore include qualitative analysis of retrieved nuggets and their role in legal reasoning. Finally, the experiments cover four Brazilian legal datasets and a limited set of embedding models, so the conclusions should be tested on additional legal collections and retrieval settings.

6. Conclusion

This paper investigated whether legal nugget extraction can improve dense retrieval over long Brazilian jurisprudential texts. The proposed pipeline extracts atomic legal theses from source documents, indexes these nuggets with embeddings, retrieves nugget-level evidence for each query, and aggregates the retrieved evidence back to document rankings.

The results support a qualified answer. Nugget retrieval improves retrieval on the two jurisprudential datasets evaluated with `text-embedding-3-small`, especially when queries are themselves close to legal theses. However, the approach is not uniformly better than full-document retrieval: it underperforms on NormasTCU and BR-TaxQA, and the embedding-model ablation shows that strong JUA-adapted retrievers remain better in the full-document setting. At the same time, nuggets improve general Qwen3 embedding models on JUA-Juris and improve MAP@10 in some JurisTCU comparisons, suggesting that granular representations can be useful when the embedding model has not already learned a strong document-level representation for the target benchmark.

Taken together, these findings suggest that legal nuggets should be viewed as a task-sensitive representation strategy rather than a universal replacement for full-document search. Their value depends on the alignment between the prompt, the legal genre, the query formulation, and the embedding model used for retrieval. Future work should study hybrid retrieval strategies that combine nugget and full-document evidence, better aggregation methods, cost-efficient nugget generation with smaller or open-source models, and qualitative evaluation of the faithfulness and legal usefulness of retrieved nuggets.

Ethics Statement

This work evaluates retrieval methods over existing legal benchmark datasets and does not introduce new personal data collection or user-facing legal decision systems. Nevertheless, legal retrieval is a high-stakes domain: retrieved documents may influence legal research, institutional work, or professional judgment. Nugget-based retrieval can make relevant legal theses easier to find, but it can also omit context, surface incomplete propositions, or overemphasize fragments that should be interpreted together with the full decision. For this reason, the method should be viewed as a research tool for retrieval and evidence discovery, not as a substitute for legal analysis by qualified professionals.

Declaration of Generative AI Use

The authors used generative AI tools to support language editing and grammar correction during the preparation of this manuscript. Generative AI tools were not used to generate experimental results, fabricate data, or perform the quantitative evaluation reported in this paper. All scientific claims, system descriptions, experimental results, and final text were reviewed, verified, and approved by the authors, who remain fully responsible for the content of the submission.

References

Arabzadeh, N. and Clarke, C. L. A. (2025). Benchmarking LLM-based relevance judgment methods. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3194–3204.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. (2020). LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904.

Dietz, L., Li, B., Liu, G., Ju, J.-H., Yang, E., Lawrie, D., Walden, W., and Mayfield, J. (2026). Incorporating Q&A nuggets into retrieval-augmented generation.

Farzi, N., Dietz, L., and Lewis, D. D. (2026). Supporting humans in evaluating AI summaries of legal depositions. In *Proceedings of the 2026 Conference on Human Information Interaction and Retrieval*, pages 350–354.

Fernandes, L. C., Ribeiro, L. d. S., de Castro, M. V. B., da Silva Pacheco, L. A., and de Oliveira Sandes, E. F. (2026). JurisTCU: a Brazilian Portuguese information retrieval dataset with query relevance judgments. *Language Resources and Evaluation*, 60(1):23.

Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446.

Júnior, J. D., Faria, A., de Oliveira, E. S., de Brito, E., Teotonio, M., Assumpção, A., Carmo, D., Lotufo, R., and Pereira, J. (2026). BR-TaxQA-R: A dataset for question answering with references for brazilian personal income tax law, including case law. In de Freitas, R. and Furtado, D., editors, *Intelligent Systems*, pages 208–222, Cham. Springer Nature Switzerland.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6769–6781.

Khattab, O. and Zaharia, M. (2020). ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 39–48.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Ma, Y., Wu, Y., Ai, Q., Liu, Y., Shao, Y., Zhang, M., and Ma, S. (2024). Incorporating structural information into legal case retrieval. *ACM Transactions on Information Systems*, 42(2):1–28.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, Cambridge.

Pereira, J., Fernandes, L., de Brito, E., Lotufo, R., and Bonifacio, L. (2026a). JUÁ - a benchmark for information retrieval in brazilian legal text collections.

Pereira, J., Lotufo, R., and Bonifacio, L. (2026b). Domain-adaptive dense retrieval for brazilian legal search.

Thakur, N., Lin, J., Havens, S., Carbin, M., Khattab, O., and Drozdov, A. (2025). FreshStack: Building realistic benchmarks for evaluating retrieval on technical documents. In *Advances in Neural Information Processing Systems*.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., and Gurevych, I. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

Voorhees, E. M. and Tice, D. M. (2000). Building a question answering test collection. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 200–207.

Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., Xie, P., Yang, A., Liu, D., Lin, J., Huang, F., and Zhou, J. (2025). Qwen3 Embedding: Advancing text embedding and reranking through foundation models.

Łajewska, W. and Balog, K. (2025). GINGER: Grounded information nugget-based generation of responses. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2723–2727.

The Impact of Language and Portuguese Intralinguistic Variation on NL2SQL

Ricardo Trainotti Rabonato , Lilian Berton

[1]Universidade Federal de São Paulo (UNIFESP)
São José dos Campos – SP – Brazil

trainotti.ricardo@unifesp.br, lberton@unifesp.br

Abstract. *Natural Language to Structured Query Language (NL2SQL) systems have advanced with large language models (LLMs), yet evaluations remain centered on English, leaving open how these systems behave across languages and language varieties. We aim to contribute to this by analyzing two state-of-the-art LLMs (LLaMA 3.3 70B and Qwen 3.6 35B, both zero-shot) on the WikiSQL benchmark in English, Brazilian Portuguese (PT-BR), and European Portuguese (PT-PT), keeping database schemas and inference configuration fixed. Translations were validated through distance metrics (cosine similarity 0.966), contrastive linguistic analysis confirming varietal features (e.g., cleft constructions 36.8× more frequent in PT-PT), and multi-model judgment (85% / 81% semantic equivalence, 95.5% PT-BR and 96.0% PT-PT inter-judge agreement). Results show a marked drop from English (47.25% EX) to Portuguese (PT-BR: 40.03%; PT-PT: 39.34%), with the PT-BR vs. PT-PT difference reaching statistical significance ($p = 0.019$). English-only successes outnumber Portuguese-only successes nearly 10:1. A replication on a second model (Qwen 3.6 35B) confirms the English-Portuguese gap but suggests that intralinguistic effects may be model-dependent. Stratified analyses show semantic complexity and quantification amplify degradation, particularly for COUNT queries. These findings point to persistent English bias in NL2SQL and measurable effects of intralinguistic variation, even between closely related varieties.*

1. Introduction

Language models have played an increasingly important role in *semantic parsing*, particularly in the conversion of natural language into SQL queries (NL2SQL) [Liu et al. 2024, Zhao et al. 2025]. Large language models (LLM) now enable complex query generation from natural language prompts in *zero-shot* or *few-shot* settings, reducing reliance on specialized architectures and expanding NL2SQL applicability in analytical and corporate systems [Katsogiannis-Meimarakis and Koutrika 2023, Fan et al. 2024]. Despite these advances, research remains strongly centered on English-language data and benchmarks such as WikiSQL and Spider [Zhong et al. 2017, Yu et al. 2018], leaving open how NL2SQL systems behave in other languages—especially those with substantial internal variation, like Portuguese.

Portuguese exhibits a well-established division between its two most widely used written varieties: Brazilian Portuguese (PT-BR) and European Portuguese (PT-PT). Although mutually intelligible, these varieties display differences in vocabulary, syntax, polysemy, lexical preferences, and register [Azevedo 2005]. Prior work indicates that linguistic differences between varieties of the same language can affect multilingual model performance, particularly when these varieties are unevenly represented in pretraining

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 317–330
October 19-22, 2026 ©2026 Association for Computational Linguistics

data [Pires et al. 2019]. However, evidence remains scarce on how such differences manifest in tasks with high semantic complexity—such as SQL query generation, which demands precise intent interpretation, entity identification, accurate column mapping, and multi-step logical reasoning [Liu et al. 2024, Zhao et al. 2025].

This gap matters because current LLMs exhibit a strong bias toward English, which dominates both pretraining corpora and evaluation benchmarks. When deploying NL2SQL for Portuguese, two distinct linguistic factors come into play: (i) cross-lingual transfer from English to Portuguese [Shi et al. 2022], and (ii) intralinguistic variation between PT-BR and PT-PT [Preda et al. 2024]. While cross-lingual transfer has received considerable attention, the impact of intralinguistic variation remains largely unexplored—despite its practical relevance for applications serving users across Brazil and Portugal, where the same system must handle both varieties.

In this work, we investigate how general-purpose LLMs perform NL2SQL when exposed to questions in English, PT-BR, and PT-PT. Our primary experiments use LLaMA 3.3 70B on the full WikiSQL benchmark [Zhong et al. 2017], with a replication on Qwen 3.6 35B to assess cross-model generalization. Questions were translated into both Portuguese varieties and validated through a multi-layered protocol combining distance metrics, contrastive linguistic analysis, and multi-model judgment. By keeping the database schema, instructions, and hyperparameters identical across languages, we isolate the impact of input language and intralinguistic variation. We evaluate performance using *Execution Accuracy* and *Logical Form Accuracy*, stratified by query complexity and aggregation type, complemented by qualitative error analysis.

The contributions of this work are threefold:

- A trilingual evaluation (EN, PT-BR, and PT-PT) of WikiSQL using a controlled NL2SQL setup with two models from different families, quantifying the impact of input language and assessing cross-model consistency.
- The first analysis of Portuguese intralinguistic variation on NL2SQL performance, identifying error patterns tied to semantic complexity and aggregation type.
- A parallel corpus of WikiSQL questions translated into PT-BR and PT-PT, that will be released for future research on NL2SQL in Portuguese.

2. Related Work

NL2SQL has evolved from early sequence-to-sequence architectures such as Seq2SQL [Zhong et al. 2017] and SQLNet [Xu et al. 2017] to specialized models with structure-aware decoding, including RAT-SQL [Wang et al. 2020] and PICARD [Scholak et al. 2021]. Standardized benchmarks like WikiSQL and Spider [Yu et al. 2018] enabled systematic comparisons across models, yet the field remains strongly centered on English.

With large language models, NL2SQL can be performed via few-shot prompting or instruction tuning, without task-specific architectures [Katsogiannis-Meimarakis and Koutrika 2023, Fan et al. 2024]. Models such as T5 [Raffel et al. 2020], Codex [Chen et al. 2021], and LLaMA [Touvron et al. 2023] generate high-quality SQL queries even for unseen databases—but they are predominantly trained on English corpora, with limited exposure to technical Portuguese.

Portuguese NLP resources exist—BERTimbau [Souza et al. 2020], Universal Dependencies treebanks [Rademaker et al. 2017], BrWaC [Wagner Filho et al. 2018]—yet focus predominantly on Brazilian Portuguese. Cross-lingual transfer studies show that multilingual models such as mBERT degrade on underrepresented language variants [Pires et al. 2019], though these analyses have been restricted to lower-level tasks like POS tagging, leaving open the question for semantically complex tasks.

Recent work has begun addressing multilingual NL2SQL [Shi et al. 2022] and Portuguese dialect identification [Preda et al. 2024]. To our knowledge, however, no prior study systematically investigates the impact of Portuguese intralinguistic variation on NL2SQL, nor provides parallel benchmarks covering PT-BR and PT-PT. This gap directly motivates the present work.

3. Method

Our experimental design isolates the impact of input language and intralinguistic variation by varying only the language of the natural language question, while holding all other factors constant: database schema, gold SQL queries, and inference parameters.

3.1. Base Benchmark

We use WikiSQL, one of the most widely adopted datasets for evaluating NL2SQL systems [Zhong et al. 2017]. WikiSQL pairs natural language questions with corresponding SQL queries over real Wikipedia tables. Each example includes a question, a structured SQL representation, and a target table, enabling both execution-based and structural evaluation. Although originally proposed in English, WikiSQL's relatively simple structure—selection, aggregation, and filtering conditions—makes it well suited for controlled comparisons across linguistic variants.

3.2. Question Translation

Starting from the original English questions in the WikiSQL development set, we constructed two translated versions: PT-BR and PT-PT. Translations were produced automatically using the same LLaMA 3.3 70B model used in the main experiment, with explicit instructions to avoid simplification and to maintain interrogative structures and lexical choices compatible with each variety. No manual post-editing was performed, nor were cultural adaptations or free reformulations introduced. Differences between PT-BR and PT-PT thus predominantly reflect linguistic preferences learned during pretraining. The goal was to produce linguistically plausible, semantically equivalent versions of the original questions, without task-specific optimization for NL2SQL.

3.2.1. Translation Validation

We assessed translation quality through three complementary layers: automatic distance metrics over the full corpus, contrastive linguistic analysis, and multi-model judgment on a stratified sample.

Distance Metrics. Comparing each PT-BR/PT-PT pair across the full corpus (N=8,421) revealed high semantic equivalence with consistent surface divergence. Mean cosine similarity computed with multilingual sentence embeddings[1] was 0.966 (SD=0.058), confirming near-perfect semantic preservation. BLEU-4 averaged 0.709 (SD=0.298) and normalized Levenshtein distance 0.104 (SD=0.113)—indicating systematic lexical and syntactic differences between variants despite shared meaning. Token count ratios averaged 0.986 (SD=0.157), ruling out systematic lengthening or shortening in either variant.

Linguistic Characterization. To verify that translations reflect authentic varietal differences rather than superficial lexical substitution, we analyzed syntactic features known to distinguish the two varieties. Table 1 summarizes the results.

Feature	PT-BR	PT-PT
Cleft constructions ("é que")	4	147
Gerund periphrases ("está correndo")	30	0
A + infinitive ("está a correr")	0	11
Enclitic pronoun placement	0	24
Proclitic pronoun placement	719	659
Article + possessive ("o meu")	7	50

Table 1. Frequency of varietal linguistic features in the PT-BR and PT-PT translations (N=8,421). The sharp contrasts (e.g., cleft constructions 36.8× more frequent in PT-PT, gerund periphrases exclusive to PT-BR) validate that translations authentically reflect each variety's syntactic preferences.

Cleft constructions—a well-known marker of European Portuguese—appear 36.8 times more frequently in PT-PT than in PT-BR. Gerund periphrases occur exclusively in PT-BR, while the a + infinitive construction occurs exclusively in PT-PT, mirroring the documented aspectual divide between the varieties. Enclisis appears only in PT-PT translations, and the rate of article + possessive sequences is over seven times higher in PT-PT (50 vs. 7). These patterns confirm that the translations respect varietal syntactic norms.

Multi-Model Quality Judgment. We further validated a stratified sample of 200 question pairs—100 where English succeeded but Portuguese failed, 50 where all languages succeeded, and 50 random—using three independent LLMs as judges (DeepSeek V4 Flash, Gemma 4 31B, Qwen 3.6 35B). Each judge assessed translations along three criteria using a standardized scale: *semantic equivalence* (YES / PARTIAL / NO), *varietal naturalness* (NATURAL / ACCEPTABLE / UNNATURAL), and *complexity preservation* (YES / PARTIAL / NO), with judges instructed to respond exclusively in JSON format. Inter-judge agreement was high: at least two of three judges concurred in 95.5% of cases for PT-BR and 96.0% for PT-PT.

The majority vote was "YES" for semantic equivalence in 85.0% of PT-BR and 81.0% of PT-PT translations, with only 6.0% and 5.5% rated "NO." Complexity preservation was near-perfect (97.0% PT-BR, 95.0% PT-PT). Varietal naturalness ratings were "NATURAL" or "ACCEPTABLE" for 67.5% of PT-BR and 72.0% of PT-PT translations.

[1] Using `paraphrase-multilingual-MiniLM-L12-v2` from the Sentence Transformers library.

Taken together, these analyses indicate that the translations are semantically faithful, exhibit authentic varietal features, and are free of systematic artifacts that could confound the NL2SQL evaluation.

3.3. Model Configuration

We use LLaMA 3.3 70B [Meta AI 2024] in a *zero-shot* setting across all conditions, without fine-tuning or task-specific training. This lets us analyze how a widely used generalist model responds to linguistic variation in the input. We stress that our goal is not to assess absolute NL2SQL quality, but to measure relative performance differences arising exclusively from the language of the question.

To ensure direct comparability, we keep the following fixed across all experiments:

- table schemas (in English, as in the original WikiSQL);
- input format;
- instruction prompt (in English, unchanged across conditions);
- decoding and inference parameters;
- number of attempts per query.

Under this design, any observed variation in performance is attributable exclusively to the language of the natural language question.

For cross-model validation, we replicate the main experiment on a stratified sample of 500 examples using Qwen 3.6 35B [Qwen Team 2026], a Mixture-of-Experts model from a different family (Alibaba). All other configuration parameters (schemas, prompts, translations, and decoding settings) remain identical.

3.4. Evaluation Protocol

Evaluation was conducted separately for each linguistic condition (EN, PT-BR, PT-PT) on the same examples and WikiSQL tables. We employ two complementary metrics:

- **Execution Accuracy (EX)**: proportion of queries whose execution result matches the gold result when both queries are run on the database, regardless of surface differences in SQL formulation. Two queries are considered equivalent if they return identical result sets, even if their SQL structures differ;
- **Logical Form Accuracy (LF)**: proportion of queries whose parsed representation exactly matches the gold query's parsed representation. Following the WikiSQL parser [Zhong et al. 2017], we convert each SQL dictionary into a `Query` object and consider a prediction correct if its parsed structure is identical to the gold structure (qp == qg), including all conditions, aggregations, and ordering.

Execution Accuracy captures functional behavior in practical scenarios, while Logical Form Accuracy provides a stricter measure of structural correspondence. Both were computed identically across conditions.

3.5. Analysis by Query Type

We stratify results by structural characteristics of the gold SQL queries: number of conditions in the `WHERE` clause, and aggregation type in the `SELECT` clause (`COUNT`, `SUM`, `AVG`, etc.). This lets us examine whether linguistic effects are uniform or intensify with query complexity.

3.6. Qualitative Error Analysis

We complement quantitative results with exploratory inspection of selected examples, focusing on cases where the English query succeeds but Portuguese versions fail, and where PT-BR and PT-PT diverge. The analysis is not exhaustive; its purpose is to illustrate recurring patterns—difficulties with quantification, multi-condition combinations, or interrogative constructions—that support the statistical trends identified in our experiments.

4. Results

Evaluation was conducted on the WikiSQL development set (8,421 examples per condition). Table 2 summarizes overall performance.

Language	EX (%)	LF (%)
English (EN)	**47.25**	28.93
Brazilian Portuguese (PT-BR)	**40.03**	22.53
European Portuguese (PT-PT)	**39.34**	22.30

Table 2. Execution (EX) and Logical Form (LF) accuracy of LLaMA 3.3 70B (zero-shot) on the WikiSQL development set, with schemas and inference parameters held constant across English, PT-BR, and PT-PT.

4.1. Overall Performance

Performance drops sharply from English to Portuguese. The model reaches 47.25% Execution Accuracy in English, falling to 40.03% for PT-BR and 39.34% for PT-PT—a gap of 7–8 percentage points attributable entirely to input language, since schemas and SQL targets remained fixed. Logical Form Accuracy follows the same pattern (28.93% EN vs. 22.53% PT-BR and 22.30% PT-PT).

4.2. Comparison between PT-BR and PT-PT

Differences between Portuguese variants are smaller but consistent: PT-BR outperforms PT-PT by 0.69 points in EX and 0.23 in LF, persisting across both evaluation dimensions.

4.3. Statistical Tests

We applied the paired McNemar test to all language pairs. Differences are highly significant for EN vs. PT-BR ($\chi^2 = 289.2$, $p < 0.001$) and EN vs. PT-PT ($\chi^2 = 341.2$, $p < 0.001$). The PT-BR vs. PT-PT difference also reaches significance ($\chi^2 = 5.51$, $p = 0.019$): intralinguistic variation has a measurable, non-random effect. Discordant pairs between Portuguese variants account for 7.0% of examples (590 of 8,421), with a slight PT-BR advantage (324 vs. 266 exclusive correct predictions).

4.4. Error Overlap Analysis

Table 3 shows how correctness patterns distribute across languages.

English-only successes (8.9%) outnumber Portuguese-only successes (1.1% and 0.9%) by nearly an order of magnitude, reinforcing that input language—not random variation—drives the gap. Among Portuguese errors, both variants share the same outcome in the vast majority of cases (4,784 both incorrect vs. 590 discordant), consistent with a common English-centric pretraining bias. The 590 discordant cases are where intralinguistic variation leaves a measurable footprint.

Pattern	Count	%
All three correct	2,806	33.3
All three incorrect	4,035	47.9
EN only correct	749	8.9
PT-BR only correct	92	1.1
PT-PT only correct	74	0.9
EN and PT-BR correct (PT-PT wrong)	232	2.8
EN and PT-PT correct (PT-BR wrong)	192	2.3
PT-BR and PT-PT correct (EN wrong)	241	2.9

Table 3. Overlap of execution correctness across English (EN), Brazilian Portuguese (PT-BR), and European Portuguese (PT-PT) for the 8,421 queries. "EN only correct" cases outnumber Portuguese-only successes nearly 10:1, indicating strong English-centric bias.

4.5. Performance by Query Type and Aggregation

Tables 4 and 5 stratify Execution Accuracy by WHERE conditions and aggregation operator.

Query Type	N	EN	PT-BR	PT-PT
Conds = 0	64	**78.12**	73.44	67.19
Conds = 1	5,755	**53.19**	45.66	45.54
Conds = 2	2,082	**33.77**	26.95	24.93
Conds $\geq$ 3	520	**31.73**	25.96	25.00

Table 4. Execution accuracy stratified by number of conditions in the `WHERE` clause (N = number of examples per class). The English-Portuguese gap widens with query complexity, showing that semantic complexity amplifies linguistic mismatch.

Performance degrades as complexity increases; while all languages suffer, the English-Portuguese gap remains evident throughout.

Aggregation	N	EN	PT-BR	PT-PT
NONE	6,017	49.04	42.94	42.70
COUNT	779	29.91	14.76	11.17
MAX	507	50.89	43.98	41.62
MIN	468	52.99	44.23	42.09
SUM	321	41.12	32.71	35.51
AVG	329	47.72	41.64	41.03

Table 5. Execution accuracy by aggregation operator in the `SELECT` clause (N = number of examples per operator). The `COUNT` operator shows the most severe degradation in Portuguese, with PT-PT accuracy (11.17%) reaching only one-third of the English baseline (29.91%).

The `COUNT` operator shows the most severe degradation: PT-PT accuracy (11.17%) is roughly one-third of EN accuracy (29.91%). Quantification operators (`SUM`, `AVG`) also amplify sensitivity to input language, particularly in European Portuguese.

4.6. Cross-Model Replication

To assess whether the observed patterns are specific to LLaMA 3.3 70B, we replicated the experiment on a stratified random sample of 500 examples using Qwen 3.6 35B

[Qwen Team 2026], a model from a different family and architecture (Mixture-of-Experts, Alibaba). We kept the same translations, schemas, prompts, and inference configuration.

Table 6 summarizes the results. Both models exhibit a clear performance drop from English to Portuguese, confirming that the English bias is not an artifact of a single model. The gap is larger for LLaMA (6.20 pp) than for Qwen (3.00 pp), but the direction is consistent.

Model	EN	PT-BR	PT-PT	Δ EN$-$PT
LLaMA 3.3 70B	44.80	40.40	38.60	6.20
Qwen 3.6 35B	51.60	48.20	48.60	3.00

Table 6. Cross-model comparison of Execution Accuracy (EX) on a stratified random sample of 500 examples. LLaMA 3.3 70B and Qwen 3.6 35B both show a consistent performance drop from English to Portuguese, confirming that the English-centric bias is not model-specific. Δ EN$-$PT denotes the accuracy gap between English and the average of PT-BR and PT-PT.

We note that LLaMA's accuracy on this 500-example sample (44.80% EN) differs slightly from the full corpus (47.25%) due to the stratified sampling, which oversampled complex queries relative to their corpus proportion.

We applied the paired McNemar test to Qwen's predictions on the 500-example sample. None of the pairwise comparisons reached statistical significance at the 5% level: EN vs. PT-BR ($\chi^2 = 3.24$, $p = 0.072$), EN vs. PT-PT ($\chi^2 = 2.42$, $p = 0.120$), and PT-BR vs. PT-PT ($\chi^2 = 0.03$, $p = 0.864$). We attribute this primarily to the reduced sample size (500 vs. 8,421 for LLaMA), which limits statistical power. The numerical trends, however, align with LLaMA's: English outperforms both Portuguese varieties, and PT-BR and PT-PT are virtually indistinguishable in Qwen's predictions (48.20% vs. 48.60%).

The intralinguistic pattern differs between models. LLaMA consistently favors PT-BR over PT-PT across all query types and aggregation operators, with a statistically significant gap on the full corpus. Qwen shows no systematic advantage for either variant: the direction varies across strata and the gap is negligible (0.4 pp). This suggests that while the English-Portuguese gap is a robust cross-model phenomenon, the subtler effects of intralinguistic variation may depend on model-specific factors such as pretraining data composition and tokenization.

For the COUNT operator, LLaMA shows the most severe degradation in Portuguese (25.0% EN vs. 8.3% for both PT-BR and PT-PT on the sample), while Qwen performs comparably across languages (50.0% EN vs. 54.2% PT-BR vs. 50.0% PT-PT). We caution that the COUNT stratum contains only 48 examples, so these figures should be interpreted as indicative rather than conclusive.

4.7. Summary

Across all analyses, input language emerges as a decisive factor in NL2SQL performance for both models tested. The English-Portuguese gap is consistent in direction and substantial in magnitude: 7–8 percentage points for LLaMA on the full corpus and 3–6 points for both models on the stratified sample, with semantic complexity and quantification operators amplifying the degradation. The intralinguistic pattern, however, proved

model-dependent. LLaMA shows a statistically significant but small PT-BR advantage ($p = 0.019$), while Qwen exhibits no systematic preference for either variant. This suggests that while English-centric bias is a robust cross-model phenomenon, the subtler effects of Portuguese intralinguistic variation may depend on specific pretraining data composition and tokenization strategies, warranting further investigation across a broader range of models.

5. Discussion

5.1. English Bias in NL2SQL

The sharp drop from English to Portuguese—under identical schemas and SQL targets—points to a pronounced English bias in these models, likely rooted in the predominance of English in pretraining data and in NL2SQL benchmarks. Since database schemas remained in English, the degradation reflects the added difficulty of mapping Portuguese questions to English schema elements: the core cross-lingual challenge. Stratified results show this is not uniform: queries with multiple conditions and quantification operators amplify the degradation, indicating that semantic complexity compounds linguistic mismatch.

5.2. Intralinguistic Variation: PT-BR vs. PT-PT

The statistically significant PT-BR vs. PT-PT difference in Llama ($p = 0.019$), calls for explanation, though we note that this pattern did not replicate in Qwen, where the two variants performed nearly identically (48.20% vs. 48.60%). For LLamA, we see two plausible factors. First, the imbalance in pretraining data representation, where PT-BR tends to be more prevalent, may give the model marginally better adaptation to its lexical and syntactic patterns. Our linguistic validation confirms that translations exhibit authentic varietal features: cleft constructions 36.8$\times$ more frequent in PT-PT, enclisis exclusive to PT-PT, gerund periphrases exclusive to PT-BR. Llama thus faces genuinely different input distributions across variants.

Second, automatic translation itself may introduce subtle artifacts. PT-BR translations received slightly higher semantic equivalence ratings (85.0% vs. 81.0%), which could contribute to the gap. We treat these as explanatory hypotheses rather than definitive conclusions, given the absence of human-translated references. The fact that Qwen does not exhibit a PT-BR advantage suggests that intralinguistic effects, unlike the English-Portuguese gap, are not uniform across models and may depend on specific pretraining data mixtures or tokenization choices.

5.3. Qualitative Error Patterns

Inspecting discordant cases reveals patterns beyond lexical differences. Identical questions across variants sometimes produced different errors—schema linking failures independent of translation quality. Semantically equivalent quantification expressions (*"pontuação total"* vs. *"total de pontos"*) triggered different aggregation operators, suggesting sensitivity to how totality is syntactically framed. Temporal expressions with opposite interpretations (*"primeiros anos"* vs. *"ano mais antigo"*) led to opposite operators (MIN vs. MAX). These cases show how minor intralinguistic differences propagate

into structurally distinct SQL queries, aligning with broader findings in semantic parsing where subtle linguistic variation produces systematic errors in column and operator identification.

The severe degradation of `COUNT` (29.91% EN vs. 14.76% PT-BR vs. 11.17% PT-PT) reflects a structural challenge absent from other operators. Whereas `MAX`, `MIN`, `SUM`, and `AVG` typically anchor to an explicit numeric column, `COUNT` must be inferred from the question's semantic intent (e.g., *"quantos jogadores"* vs. *"qual o número de"* vs. *"conte"*), offering no direct lexical anchor to guide schema linking. This makes the operator uniquely dependent on cross-lingual parsing of quantification intent—precisely where the model's English-centric pretraining bias is most consequential. The pattern aligns with the sensitivity to syntactic framing observed across Portuguese variants, where semantically equivalent expressions triggered different aggregation operators.

5.4. Implications

These findings underscore the need for multilingual and multi-varietal evaluation in NL2SQL. Relying solely on English benchmarks risks overestimating system robustness. For practitioners, even small differences between language varieties can produce measurable degradation, which is a concern in business intelligence and data analysis applications serving diverse Lusophone users. The cross-model replication further suggests that while English-centric bias is a shared limitation, intralinguistic sensitivity varies by model, complicating one-size-fits-all solutions.

6. Conclusion

We presented the first systematic evaluation of input language and Portuguese intralinguistic variation on NL2SQL performance. Using translated versions of WikiSQL, we compared two LLMs (LLaMA 3.3 70B and Qwen 3.6 35B, both zero-shot) across English, PT-BR, and PT-PT, keeping schemas, SQL targets, and configuration fixed.

Performance drops markedly from English (47.25% EX) to Portuguese (PT-BR: 40.03%; PT-PT: 39.34%), with the PT-BR vs. PT-PT difference reaching statistical significance ($p = 0.019$). English-only successes outnumber Portuguese-only successes nearly 10:1, and 7.0% of examples show discordant outcomes between Portuguese variants. Stratified analyses reveal that semantic complexity and quantification amplify these effects: `COUNT` queries exhibit the most severe degradation in Portuguese. A replication with Qwen 3.6 35B on a stratified sample confirmed the cross-model robustness of the English-Portuguese gap, while intralinguistic patterns varied, suggesting that the latter are sensitive to model-specific factors such as pretraining data composition.

A multi-layered translation validation with distance metrics, contrastive linguistic analysis, and multi-model judgment confirmed that the translations are semantically faithful (cosine similarity 0.966, semantic equivalence 85.0% PT-BR and 81.0% PT-PT) while exhibiting authentic varietal features (cleft constructions 36.8$\times$ more frequent in PT-PT, enclisis exclusive to PT-PT). The parallel PT-BR/PT-PT WikiSQL corpus will be released for future research.

These findings demonstrate that despite advances in multilingual LLMs, English bias remains a significant challenge for NL2SQL, and intralinguistic variation, while sec-

ondary to language change, produces measurable effects that can reach statistical significance and should not be overlooked in multilingual deployments.

7. Limitations

Four aspects of this study constrain the scope of our conclusions. First, evaluation was restricted to WikiSQL, whose relatively simple queries may not represent complex NL2SQL scenarios involving multi-table joins or nested queries. Second, translations were obtained automatically. Although we conducted extensive validation—distance metrics, contrastive linguistic analysis, and multi-model judgment—these methods do not replace human evaluation, and subtle artifacts may account for part of the PT-BR/PT-PT differences. Third, while both LLaMA 3.3 70B and Qwen 3.6 35B exhibited consistent English-Portuguese gaps, intralinguistic patterns differed between them. Fourth, absolute performance remains below the state of the art for fine-tuned NL2SQL models. Across all dimensions, future work should pursue human-validated parallel corpora, evaluations on complex benchmarks such as Spider, and replication across a broader set of model families.

8. Ethical Considerations

This work uses publicly available datasets (WikiSQL) and does not involve human subjects, proprietary data, or deployment in production systems. The translations were generated automatically and validated through automated methods; no human annotators were employed. We plan to release the parallel PT-BR/PT-PT corpus to foster research on Portuguese NLP, with no restrictions beyond attribution. We do not foresee direct harmful applications of this work, though we acknowledge that NL2SQL systems may perpetuate biases present in pretraining data, including asymmetries between language varieties.

9. Generative AI Use Declaration

Large language models were used as experimental tools in this study: for translating the WikiSQL benchmark into PT-BR and PT-PT (LLaMA 3.3 70B), for generating SQL queries from natural language questions (LLaMA 3.3 70B and Qwen 3.6 35B), and for validating translation quality through multi-model judgment (DeepSeek V4 Flash, Gemma 4 31B, Qwen 3.6 35B). Additionally, generative AI was used in the preparation of approximately 25% of the manuscript, primarily for polishing and condensing text drafted by the authors. The authors reviewed and edited all AI-generated content and remain fully responsible for the final manuscript.

Acknowledgments

This study was financed, in part, by the São Paulo Research Foundation (FAPESP), Brasil. Process Number 2025/03245-1.

References

Azevedo, M. M. (2005). *Portuguese: A linguistic introduction*. Cambridge University Press.

Chen, M. et al. (2021). Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Fan, J., Gu, Z., Zhang, S., Zhang, Y., Chen, Z., Cao, L., Li, G., Madden, S., Du, X., and Tang, N. (2024). Combining small language models and large language models for zero-shot nl2sql. *Proceedings of the VLDB Endowment*, 17(11):2750–2763.

Katsogiannis-Meimarakis, G. and Koutrika, G. (2023). A survey on deep learning approaches for text-to-sql. *The VLDB Journal*, 32(4):905–936.

Liu, X., Shen, S., Li, B., Ma, P., Jiang, R., Zhang, Y., Fan, J., Li, G., Tang, N., and Luo, Y. (2024). A survey of nl2sql with large language models: Where are we, and where are we going? *arXiv preprint arXiv:2408.05109*.

Meta AI (2024). Llama 3.3 70b model card. `https://github.com/meta-llama/llama-models/blob/main/models/llama3_3/MODEL_CARD.md`. Accessed: 2025-12-10.

Pires, T., Schlinger, E., and Garrette, D. (2019). How multilingual is multilingual bert? In *Proceedings of ACL*.

Preda, D., Osório, T., and Cardoso, H. L. (2024). Across the atlantic: Distinguishing between european and brazilian portuguese dialects. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 1*, pages 353–363.

Qwen Team (2026). Qwen3.6-35B-A3B: Agentic coding power, now open to all.

Rademaker, A., Chalub, F., Real, L., Freitas, C., Bick, E., and de Paiva, V. (2017). Universal Dependencies for Portuguese. In Montemagni, S. and Nivre, J., editors, *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 197–206, Pisa, Italy. Linköping University Electronic Press.

Raffel, C. et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Scholak, T., Schucher, N., and Bahdanau, D. (2021). Picard: Parsing incrementally for constrained auto-regressive decoding from language models.

Shi, P., Zhang, R., Bai, H., and Lin, J. (2022). Xricl: Cross-lingual retrieval-augmented in-context learning for cross-lingual text-to-sql semantic parsing. *arXiv preprint arXiv:2210.13693*.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Touvron, H. et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wagner Filho, J. A., Wilkens, R., Idiart, M., and Villavicencio, A. (2018). The brWaC corpus: A new open resource for Brazilian Portuguese. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Wang, B. et al. (2020). Rat-sql: Relation-aware schema encoding and linking for text-to-sql parsers. In *Proceedings of ACL*, pages 7567–7578.

Xu, X., Liu, C., and Song, D. (2017). Sqlnet: Generating structured queries from natural language without reinforcement learning.

Yu, T., Zhang, R., Yang, K., et al. (2018). Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. In *Proceedings of EMNLP*, pages 3911–3921.

Zhao, P., Shao, M., Cui, J., and Zhou, H. (2025). A survey of nl2sql research: From foundations to frontiers. In *2025 2nd International Conference on Intelligent Perception and Pattern Recognition (IPPR)*, pages 396–402. IEEE.

Zhong, V., Xiong, C., and Socher, R. (2017). Seq2sql: Generating structured queries from natural language using reinforcement learning.

Appendix: Prompts Used in Experiments

Translation Prompt (LLaMA 3.3 70B)

System:

> You are a professional translator. Your task is to translate English questions into [Brazilian Portuguese (PT-BR) / European Portuguese (PT-PT)] with high fidelity. Preserve the exact meaning. Do NOT add explanations, comments, markdown, or formatting. Do NOT change numbers, years, names, or entities. Translate as literally as possible while keeping natural grammar. Output ONLY the translated text.
>
> **User:**
>
> Translate the following English question into [Brazilian Portuguese (PT-BR) / European Portuguese (PT-PT)]: "[QUESTION]" Return ONLY the translated text.

NL2SQL Prompt (LLaMA 3.3 70B and Qwen 3.6 35B)

System:

> You are an NL2SQL model specialized in the WikiSQL dataset. Convert a natural language question and a table schema into a structured query in WikiSQL JSON format. Output ONLY a JSON object with fields "sel" (column index), "agg" (0=None, 1=MAX, 2=MIN, 3=COUNT, 4=SUM, 5=AVG), and "conds" (list of [col_idx, op_idx, value] where op_idx: 0==, 1=¿, 2=¡, 3=OP). Return ONLY the JSON, no explanation.
>
> **User:**
>
> Natural language question: [QUESTION]
> Table schema (column indices and names): - [0] col_name_0 - [1] col_name_1 ...
> Some example rows: - row 0: col_name_0=value, col_name_1=value, ...
> Produce ONLY the JSON with fields "sel", "agg", and "conds".

Judge Prompt (DeepSeek V4 Flash, Gemma 4 31B, Qwen 3.6 35B)

System:

You are an expert linguist evaluating translations from English to Portuguese. Assess the translation along three criteria. Answer ONLY with a valid JSON object.

User:

Evaluate the following translation:
ORIGINAL ENGLISH: "[QUESTION_EN]"
TRANSLATION ([PT-BR / PT-PT]): "[QUESTION_PT]"
Criteria: 1. SEMANTIC EQUIVALENCE: Does the translation preserve the exact meaning? (YES / PARTIAL / NO) 2. VARIETAL NATURALNESS: Does it sound natural for a native speaker of [PT-BR / PT-PT]? (NATURAL / ACCEPTABLE / UNNATURAL) 3. COMPLEXITY PRESERVATION: Are interrogative structures, negations, and quantifiers preserved? (YES / PARTIAL / NO)
Return ONLY: {"semantic_equivalence": "...", "varietal_naturalness": "...", "complexity_preservation": "..."}

Como Nossos Pais: Princípios Clássicos para a Avaliação de Sistemas de IA

Livy Real[1,2], **Altigran Soares da Silva**[2]

[1] Instituto Kunumi, Belo Horizonte, MG, Brasil
[2]Universidade Federal do Amazonas, Manaus, AM, Brasil
livy@kunumi.com, alti@icomp.ufam.edu.br

Abstract. *This work revisits epistemological and methodological foundations from several scientific fields, such as psychology, corpus linguistics, sociology, and neuroscience, in order to establish basic principles for the evaluation of systems based on Natural Language Processing and Artificial Intelligence. We present in detail the state of the art of evaluation in the field, discussing the reasons that lead to the need for automating system evaluation (the LLM-as-a-Judge paradigm). We introduce Five Basic Methodological Principles that should underpin the evaluation of AI-based systems, grounded both in theories and experiments from disciplines that have studied human perception for decades, as well as in empirical evidence from NLP itself.*

Resumo. *Este trabalho retoma bases epistemológicas e metodológicas de diversas áreas da ciência, como psicologia, linguística de corpus, sociologia e neurociência, para fundamentar princípios básicos da avaliação de sistemas baseados em Processamento de Linguagem Natural e Inteligência Artificial. Apresentamos em detalhe o estado da arte da avaliação na área, discutindo os motivos que nos levam à necessidade de automatizar a avaliação de sistemas (paradigma LLM-as-a-Judge). Introduzimos Cinco Princípios Metodológicos básicos que devem fundamentar a avaliação de sistemas baseados em IA, fundamentados tanto na teoria e experimentos destas áreas que já pesquisam a percepção humana há décadas quanto em evidências empíricas do próprio PLN.*

1. Introdução

Recentes transformações na área de Processamento de Linguagem Natural (PLN) e Inteligência Artificial (IA) passam a evidenciar a necessidade de avaliarmos os sistemas gerativos de forma mais ampla e efetiva. Ainda que motivada pela explosão da IA Gerativa (Generative AI, GenAI), a avaliação de sistemas baseados em IA e em PLN não é um tema novo. A dificuldade de avaliar as saídas de modelos gerativos [Novikova et al. 2017, Asli et al. 2020, Blagec et al. 2022], a compreensão de que muitas métricas clássicas, como BLEU e ROUGE, não eram as mais adequadas [Callison-Burch et al. 2006, Reiter and Belz 2009], a percepção de que estávamos simplificando tarefas de PLN como forma de aproximá-las de uma tarefa real [Kalouli et al. 2017, Marasović 2018], são temas conhecidos da comunidade científica. Apesar de conhecer as dificuldades da avaliação de sistemas baseados em IA, a comunidade científica não alterou o paradigma de avaliação na última década. Hoje, tarefas tradicionalmente delimitadas passaram a operar em cenários de 'mundo aberto', o potencial de geração de dados é virtualmente infinito e surgem os chamados 'comportamentos emergentes', fatores que intensificam a dificuldade de decidir **o que** avaliar.

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 331–347
October 19-22, 2026 ©2026 Association for Computational Linguistics

Nesse contexto, a avaliação de sistemas baseados em IA não pode ser tratada apenas como um problema técnico de métricas ou automação, mas como um problema metodológico e epistemológico de captura do julgamento humano. Exploramos a ausência de princípios metodológicos consolidados para operacionalizar avaliações confiáveis, reproduzíveis e alinhadas às perspectivas humanas. Como contribuição, este trabalho: (i) sistematiza desafios atuais (2026) da avaliação em PLN/IA; (ii) articula fundamentos teóricos oriundos de áreas como psicologia cognitiva, linguística de corpus, sociologia e neurociência para discutir avaliação de sistemas de IA; (iii) propõe cinco princípios metodológicos básicos, apoiados na metodologia existente de anotação de dados, para a construção de *pipelines* de avaliação humana e/ou automatizada (Princípios da *Independência, Diversidade, Confiabilidade, Simplicidade e Clareza*); (iv) fundamenta estes Princípios com exemplos práticos da literatura em PLN/IA; e (v) discute como esses princípios podem orientar o desenvolvimento de avaliações mais robustas e alinhadas ao comportamento humano em cenários reais.

Diferentemente de trabalhos focados exclusivamente em métricas, *benchmarks* ou arquiteturas de avaliação automática, nosso objetivo é propor uma perspectiva teórica e metodológica integrada para avaliação de sistemas baseados em IA, centrada na operacionalização confiável do julgamento humano (ou *human-like*). Esse trabalho deve ser lido antes como um artigo de posicionamento (*position paper*) do que como um artigo experimental. O posicionamento investigado aqui baseia-se na sólida metodologia de anotação de *corpus* [Bird and Liberman 2001, Voutilainen 2012] e em sua relação direta com as dificuldades atuais de avaliação. Os princípios metodológicos da área, por sua vez, são diretamente correlacionados com teorias da captura de julgamento humano de outras áreas da ciência, como sociologia e psicologia.

2. Avaliação Automática e Avaliação Humana

Nos últimos anos, assistimos ao *boom* do paradigma *LLM-as-a-Judge* [Lin et al. 2021, Thakur et al. 2024]. É fácil compreender esse movimento à medida que observamos a quantidade de conteúdo que temos gerado e que deveríamos avaliar. Pode-se dizer sem exageros que temos capacidade para gerar infinitos novos conteúdos, já que não existe nenhum limitador ou regulação em relação a isso. Como e quando avaliar todo esse conteúdo são questões que permanecem abertas. Considerando a escala da adoção de sistemas baseados em IA por todos os setores (público, privado, corporativo) e por sujeitos individuais, é necessário que tenhamos uma forma de avaliação em escala. A avaliação em escala é também obrigatoriamente automática[1]. Em última análise, não há humanos suficientes, suficientemente rápidos, bem treinados, bem remunerados, com a expertise ideal, para proporcionar segurança no uso de produtos baseados em IA.

Dizer que a avaliação precisa ser automatizada não é dizer que não precisamos de humanos na avaliação [Wu et al. 2022]. Sem dúvidas, precisamos de humanos na avaliação e uma avaliação só será considerada *ground truth* se for conduzida por humanos[2]. Note que uma avaliação humana, um *ground truth*, não necessariamente é a melhor ou é perfeita, livre de erros [Foody 2024]. A perfeição não vem dos humanos. O que vem

[1]Ou, ao menos, uma parte dela deve ser automatizável.

[2]Há ainda tarefas sem tanta subjetividade, como *coding* e problemas matemáticos, nas quais, eventualmente, outras formas de avaliação podem ser tão válidas quanto o julgamento humana para tarefas que envolvam linguagem natural.

do humano é a veracidade, é o compromisso (*commitment*) de que um ser humano, natural, com suas capacidades intelectuais em funcionamento, analisou um dado artificial. Um *ground truth* não é necessariamente perfeito, mas é necessariamente humano. Ele é a fonte da verdade, a representação do que humanos daquela determinada conjuntura histórica entendiam como a verdade.

Além de termos a responsabilidade humana na criação de um *ground truth*, os avaliadores humanos também estão presentes em distintos momentos desses ciclos. Por exemplo, a regulamentação brasileira em vigor, seguindo a Resolução do Conselho Nacional de Justiça (CNJ) número 615/2025 [BRASIL. Conselho Nacional de Justiça 2025], obriga que o conteúdo gerado por IA no Judiciário seja revisado e validado por um ser humano, garantindo que a responsabilidade final por um conteúdo seja sempre de um juiz ou servidor. Em atividades da indústria, é comum realizar avaliações humanas em alguns pontos do desenvolvimento de um produto e uma grande avaliação humana antes de cada novo lançamento [Artemova et al. 2025]. Recentemente, também discute-se a necessidade da perspectiva humana no que diz respeito ao 'alinhamento com humanos' [Zhao et al. 2026]: a necessidade de que as IA estejam alinhadas a princípios éticos que protejam a vida humana.

No que diz respeito ao PLN, às avaliações textuais às quais estamos habituados, o que é a introdução do alinhamento humano? É justamente entendermos o que o humano (bem-intencionado) faria ou desejaria em determinada situação e considerá-lo como *ground truth*. O que **não** é alinhamento com o humano? É um ser humano simplesmente auditar uma resposta de LLM e dizer que ela não está errada. Um humano não encontrar erros em uma resposta gerada não é a mesma coisa de garantir alinhamento humano. Um assistente automático de saúde, por exemplo, não deve ser avaliado por não gerar erros factuais, mas sim, por reproduzir o que um médico/profissional de saúde faria em uma dada situação, suas estratégias conversacionais, escolhas de registro linguístico, vocabulário técnico, entre outros. Ao passo que a avaliação automatizada está se tornando um simples e rápido *check-list* que observa se uma LLM errou ou não, estamos perdendo o alinhamento com humanos. Estamos nos alinhando às máquinas e assumindo que outro humano que leia isso aceitará esse conteúdo como **possível** ou **plausível** [Jacovi and Goldberg 2020]. O conteúdo não reflete necessariamente o que um humano especialista faria, ele apenas não possui erros factuais. Aí está a questão imediata e prática do alinhamento com humanos para o PLN. A avaliação é onde conseguimos garantir que as máquinas estão alinhadas ao que um humano faria ou deseja que ela faça, e não que humanos estão gradativamente se alinhando ao conteúdo gerado pelas máquinas.

Considerando o irrevogável papel dos humanos, seja criando o *ground truth* de uma tarefa, seja fundamentando os princípios de alinhamento humano, seja garantindo a qualidade de produto, precisamos também considerar **como** operacionalizar essas avaliações. Se a maior parte da avaliação dos conteúdos gerados será feita por máquinas, não podemos deixar essa operacionalização ser feita de maneira que não reflita o que os humanos (especialistas, se necessário) fariam.

2.1. O Paradigma *LLM-as-a-Judge*

Como operacionalizar essa forma de avaliação automática é um dos temas mais relevantes da área atualmente. O paradigma conhecido como *LLM-as-a-Judge* [Thakur et al. 2024]

tem recebido bastante atenção e a pesquisa tem se desdobrado nesse sentido. Temos inúmeros *frameworks* científicos e da indústria já disponíveis [Zheng et al. 2023, Confident AI 2024, Lee et al. 2025, Li et al. 2025b] para uso. Estudos discutem desde estratégias estruturais, como o uso de várias LLMs como juízes, conhecido como *LLM-as-a-Jury* [Qian et al. 2026], até soluções técnicas finas, como o uso de técnicas de zero-, one- ou few-shot prompting nesse contexto [Raina et al. 2024, Jang and Silavong 2025] e o uso de LLMs juízes em tarefas multimodais [Chen et al. 2024]. Também foram propostas investigações sobre métricas gerais a serem avaliadas, como *corretude* ou *completude*, referências claras aos conhecidos 'acurácia' e 'recall' [Wei et al. 2025].

À medida que se percebeu que a automação da avaliação destas tarefas não significa simplesmente replicar as métricas antigas usando modelos de linguagem [Gao et al. 2025], a própria estrutura da avaliação passou a ser investigada. Aqui, temos trabalhos que envolvem pensar na avaliação como um problema de ranqueamento [Li et al. 2025a], como uma constante escolha entre duas possibilidades (*pairwise comparison*) [Thakur et al. 2024] ou como um esquema de comparação iterativa em estilo de 'torneio', no qual múltiplas respostas são confrontadas sequencialmente até a determinação de um ranking final [Sandan et al. 2025].

Abriu-se também espaço para fazer a avaliação considerando os novos desafios que a tecnologia permite. *Groundedness*, hoje uma grande questão, no passado inexistente, torna-se mais uma variável de avaliação [Trautmann et al. 2024]. Vieses (algo que sempre existiu, mas estava menos em evidência, talvez pelos usos mais restritos de nossos modelos 'antigos', menos 'mundo aberto') passam a ser também investigados e surgem repentinamente, talvez repentinamente demais, inúmeros datasets para controle de vieses [Pagano et al. 2023]. Reforça-se nesse contexto a avaliação de temas como nocividade (*harmfulness*) e toxicidade [Hartvigsen et al. 2022].

Porém, para qualquer pessoa interessada em fazer uma avaliação real e efetiva do seu próprio sistema, com suas particularidades, essas questões gerais são pistas ou indícios, mas não avaliam exatamente o(s) ponto(s) que deve(m) ser avaliado(s). Todos esses trabalhos, especialmente os grandes datasets especializados, são muito úteis, sobretudo para a comparação de grandes modelos de linguagem, que hoje são lançados quase diariamente. Ter essas ferramentas de avaliação, *leaderboards* que tentam mensurar quase tudo, ao mesmo tempo, ajuda muito a ranquear esses modelos, a identificar como está o estado da arte e deveriam nos guiar nas escolhas de quais modelos utilizar e por quê. No entanto, ao avaliar um dado sistema, deve-se considerar o que aquele domínio e aquela tarefa especificamente precisam para serem considerados adequados em seu contexto de uso. O trabalho de [Marinho et al. 2026], por exemplo, é a tentativa de criar métricas específicas para o domínio jurídico. Um critério geral de corretude não é mais suficiente; precisamos avaliar a qualidade de cada um dos aspectos que importam nesse sistema e em seu uso real.

3. A Avaliação Ideal

Vimos que (i) o momento tecnológico impõe novas formas de avaliação caso desejemos ter controle do que produzimos com IA; (ii) não é possível escalar a avaliação sem automatizá-la, ao menos em parte; (iii) cada sistema precisará de uma avaliação diferente, não há métricas universais a serem usadas para qualquer tarefa.

Ainda que para as tarefas de mundo aberto nas quais estamos interessados atualmente não haja métricas únicas, **como** capturar o julgamento de um sujeito tem sido investigado por diferentes áreas nas últimas décadas.

O próposito deste trabalho é retomar algumas práticas tradicionais, estabelecidas por outras áreas de pesquisa, que não o PLN, que guiam como se deve coletar os dados para capturar a inteligência/compreensão humana, especialmente os dados de avaliação, mas não só. A pesquisa sobre a inteligência já existe, nós somos a parte que automatiza, artificializa essa inteligência.

Para isso, apresentamos boas práticas metodológicas de avaliação baseadas em disciplinas como linguística de corpus, psicolinguística, neurociência e historiografia. Nós, da IA, precisamos aprender a avaliar a inteligência dos nossos sistemas de inteligência artificial. Já há muito conhecimento no que diz respeito a avaliar a inteligência em si. Devemos nos apropriar desse conhecimento. Essas recomendações não são exaustivas, mas trazem princípios básicos para assegurar uma avaliação justa, seja feita por humanos, seja por LLMs. A fundamentação desses princípios foi realizada a partir de leitura crítica e orientada da literatura em anotação de dados e artigos seminais de áreas correlatas, não representando revisão sistemática de nenhuma das áreas citadas.

4. Metodologia Básica para Avaliação de Sistemas Baseados em IA

Tentaremos abordar da forma mais didática possível princípios básicos consolidados em diversas áreas do conhecimento no que diz respeito à avaliação e metodologias para a captura isenta do julgamento humano. Exemplificamos a adoção de cada um desses princípios com base na literatura recente da área de PLN. Discutimos na seção 6, como estes princípios podem ser empiricamente validados.

4.1. Princípio da Independência

O Princípio da Independência diz respeito a cada juiz conduzir seu processo avaliativo independentemente dos demais; em termos práticos, trata-se de **avaliação às cegas**.

Nenhum dos avaliadores (sejam humanos ou LLMs) deve ver a resposta de outro avaliador [Wojcieszak and Price 2009] ou ter acesso à fonte do material avaliado (por exemplo, qual modelo gerou determinada resposta). Isso evita o Viés de Falso Consenso (*False Consensus Bias*) [Ross et al. 1977], que indica a tendência de juízes concordarem com outros ou com a resposta que acreditam que a maioria irá preferir. O trabalho de [Choi et al. 2025] investiga o Viés de Falso Consenso também entre LLMs-juízes.

O Princípio da Independência significa, por exemplo, que avaliadores humanos não devem ver as respostas de LLMs-juízes e concordar ou não com respostas já estabelecidas. Isso evita o *Acquiescence Bias*, conhecido na psicologia desde a década de 1960 [Messick and Jackson 1961], que caracteriza a tendência de seres humanos responderem positivamente a uma pesquisa ou avaliação. Evita também o Efeito de Ancoragem (*Anchoring Effect*), conhecido na psicologia cognitiva desde a década de 1970 [Tversky and Kahneman 1974], que caracteriza o comportamento humano por preferir julgamentos que retomam ('ancoram') uma informação inicialmente recebida. A percepção do Efeito de Ancoragem teve implicações práticas em áreas como economia, direito, design, ciência política, entre outras.

Considerando LLMs, o Princípio da Independência evita o Viés de Acordo (*Agreement Bias*), a tendência de uma LLM concordar com uma resposta previamente estipulada [Jain et al. 2025, Thakur et al. 2024]. Ajuda também a mitigar o Viés de Simpatia (*Sycophancy Bias*), a tendência de LLMs priorizarem agradar o usuário em detrimento de raciocínio independente [Fanous et al. 2025].

Tem sido prática recorrente humanos avaliadores partirem de respostas de LLMs-juízes, concordando ou não com etiquetas ou análises previamente apresentadas. O custo mental e psicológico de discordar de algo, especialmente se o avaliador humano deve explicar por que discorda, é muito maior do que o de simplesmente concordar [Gilbert 1991]. O juiz humano tende a discordar apenas quando a resposta da LLM for ilógica. Dessa forma, muitas avaliações de sistemas de PLN recentes não são confiáveis por priorizarem economizar tempo em vez de tomar o julgamento humano *a priori* como o verdadeiro *ground truth*. Esse tipo de metodologia exemplifica o alinhamento à máquina discutido anteriormente e não é confiável. Outros conceitos, como a Pressão dos Pares (*Peer Pressure*) [Durkin 1996], da sociologia, e a Conformidade (*Conformity*) [Beloff 1958], da psicologia social, são outras formas de vieses que devem ser mitigadas pela metodologia de avaliação. Curiosamente, já temos essa prática estabelecida nos processos de revisão por pares aos quais submetemos nossos trabalhos científicos.

O trabalho de [Choi et al. 2025] reproduz o experimento seminal na psicologia de [Ross et al. 1977] procurando entender se LLMs exibem o Viés de Falso Consenso, isto é, tendem a escolher a resposta que acreditam que a maioria escolheria. Os autores testam quatro modelos (OpenAI GPT-4, Anthropic Claude 3 Opus, Meta LLaMA-2 70B, Mistral AI Mixtral 8x7B) avaliando a resposta das LLMs e a resposta que as LLMs acreditam que será escolhida pela maioria. Todos os modelos possuem o viés. Para tentar mitigá-lo, os pesquisadores testam técnicas de *prompting*, oferecendo diferentes tipos de informação às LLMs relacionadas à tarefa. Os experimentos demonstram que oferecer no prompt informações em oposição aos julgamentos iniciais reduz o Viés de Falso Consenso.

[Jain et al. 2025] mede o Viés de Simpatia de LLMs-juízes ao avaliar códigos Python de alunos. O experimento, bastante robusto, conta com 66 submissões de código Python, 69 tarefas diferentes e 14 LLMs-juízes. Os modelos avaliadores devem avaliar o código recebido, com e sem uma avaliação prévia, que pode estar correta ou incorreta. Em números gerais, as LLMs-juízes conseguem capturar 60-80% dos erros de código quando não recebem avaliação prévia, porém quando recebem a informação prévia errada, esta taxa cai para 25%. Os experimentos de [Jain et al. 2025] conseguem quantificar quanto o julgamento das LLMs-juízes é afetado quando estas têm acesso a uma resposta, mesmo que errada, evidenciando fortemente a necessidade do Princípio da Independência.

4.2. Princípio da Diversidade

O Princípio da Diversidade diz respeito à tentativa de capturar compreensões e julgamentos de avaliadores com diferentes horizontes socioculturais e *backgrounds*. Na prática, trata-se de ter **múltiplos avaliadores** para uma tarefa.

Amplamente difundida na linguística de corpus, outra prática essencial para avaliação de sistemas de PLN/IA é ter pelo menos dois juízes diferentes na avaliação de um mesmo dado [Meyer 2002, McEnery and Hardie 2012]. Se há discordância entre

eles, um terceiro juiz deve ser acessado e contabiliza-se a maioria[3]. Em tarefas muito complexas, com muitas nuances, até 5 avaliadores (ou mais) podem participar de uma mesma tarefa. Por exemplo, em tarefas que envolvem semântica e pragmática, projetos com recursos para anotação, como o corpus SICK [Marelli et al. 2014] e o conhecido SNLI [Bowman et al. 2015], possuem no mínimo 5 anotadores humanos por instância e reportam não apenas o voto majoritário, mas todos os julgamentos realizados [Frenda et al. 2024].

Embora essa boa prática seja estabelecida na linguística de corpus e linguística descritiva considerando sujeitos humanos, para LLMs a prática também se sustenta [Verga et al. 2024], sendo um avanço recente da área de avaliação usar um júri e não apenas um juiz. [Verga et al. 2024] investiga o impacto de usar múltiplos LLMs-juízes em duas tarefas, avaliação de perguntas e respostas e avaliação de chatbots. Partindo de datasets conhecidos, como o TriviaQA e o ChatBot Arena, os dados foram reanotados por humanos altamente qualificados e o acordo das LLMs-juízes foi medido em relação a esse *ground truth*. Os autores investigaram LLMs de três famílias distintas para uso individual e conjunto. Nos cenários testados, o uso de LLMs em conjunto aumenta o alinhamento entre a avaliação automática e a humana, chegando a aumentar em 10 pontos a correlação de Pearson para o benchmark ChatBotArena Hard. Os autores sugerem que essa melhoria se dá pela mitigação tanto de vieses dos modelos, quanto de inconsistências nas execuções de um mesmo modelo.

Considerando os vieses intrínsecos dos modelos, [Zhu et al. 2026] propõem uma abordagem cíclica, com um rodízio entre as LLMs-juízes. Usando o MT-Bench e 5 LLMs-juízes (Qwen 2.5 7B Instruct, Llama 3.3 70B Instruct, GPT-5.2, Gemini 3 Flash e Claude Sonnet 4.6), os autores realizam uma decomposição da variância das pontuações atribuídas, mostrando que o efeito do juiz é a principal fonte de variabilidade, respondendo por mais de 90% da variância total observada no regime de avaliação com um único juiz escolhido aleatoriamente. A variância agregada cai para 27-40% com a abordagem de rodízio de juízes (cíclica) e, ao utilizar um júri com todos os LLMs-juízes testados, ela desaparece. Obviamente, o custo da avaliação pode se tornar proibitivo no cenário com todos os juízes, logo, os autores defendem o rodízio cíclico como um compromisso eficiente entre custo e confiabilidade na avaliação baseada em LLMs.

Ter múltiplos avaliadores de uma mesma instância, então, ajuda a mitigar vieses: sejam eles culturais, individuais (de modelo ou sujeitos humanos), além de reduzir erros de anotação e alucinações. No caso de avaliadores humanos, erros são esperados em qualquer tarefa, porém são mais pronunciados quando os avaliadores estão sob pressão, seja de tempo, seja financeira, seja por condições de trabalho insalubres [Snow et al. 2008, Fort et al. 2011, Martin et al. 2014].

4.3. Princípio da Confiabilidade

O Princípio da Confiabilidade considera a necessidade de mensurar quão efetivo e confiável é o próprio processo de avaliação. Na prática, trata-se de medir o **Acordo Inter-Anotadores enquanto** a tarefa de avaliação está sendo realizada.

Para garantir qualidade e consistência na avaliação, deve-se medir regularmente

[3]Há também uma vasta literatura, já em PLN, sobre a melhor forma de computar esse desacordo e sobre como os sistemas de IA deveriam tratá-los [Pavlick and Kwiatkowski 2019, Plank 2022]

o Acordo Inter-Anotadores (*inter-annotator agreement*, IAA) entre juízes humanos e/ou LLMs [Freitas 2024]. Uma pequena parte dos dados é avaliada por todos os avaliadores na mesma tarefa; isso garante confiabilidade e reprodutibilidade no processo, assegurando que todos os envolvidos estão aptos e compreenderam a tarefa. Métricas como Cohen's Kappa, Krippendorff's Alpha, Kendall's Tau ou Scott's Pi devem ser utilizadas conforme o tipo de tarefa [Artstein and Poesio 2008]. A porcentagem simples de acordo deve ser evitada, pois desconsidera o acordo ao acaso e o viés do anotador. Valores baixos de acordo indicam a necessidade de revisar critérios, instruções e o próprio processo de julgamento. A avaliação só deve ser considerada válida quando apresentar níveis de confiabilidade compatíveis com o padrão mínimo estabelecido.

[Carletta 1996] argumenta que o acordo não deve ser tratado como mero detalhe estatístico, mas como evidência de confiabilidade experimental. Estudos posteriores mostram que baixos níveis de acordo refletem problemas estruturais da própria tarefa, incluindo ambiguidades nas diretrizes, subjetividade excessiva, categorias mal definidas ou treinamento insuficiente de anotadores [Bayerl and Paul 2011]. Nesse contexto, um acordo baixo não deve ser interpretado como falha humana, mas como sinal metodológico relevante sobre a qualidade da avaliação.

Mais recentemente, [Abercrombie et al. 2025] argumentam que medidas de acordo também funcionam como mecanismos de controle de estabilidade temporal, mostrando que anotadores humanos podem apresentar inconsistências significativas ao longo do tempo mesmo em tarefas aparentemente simples. Os autores observam variações próximas de 25% em diferentes tarefas de PLN, sugerindo que a confiabilidade avaliativa precisa ser continuamente monitorada e não apenas assumida ou *a posteriori*.

Trabalhos consolidados em LLM-as-a-Judge que comparam o alinhamento de modelos juízes entre si ou com avaliadores humanos reportam suas métricas de acordo [Zheng et al. 2023, Thakur et al. 2024, Verga et al. 2024, Lee et al. 2025, Qian et al. 2026].

4.4. Princípio da Simplicidade

O Princípio da Simplicidade ressalta a necessidade de oferecermos as tarefas mais simples possíveis para o avaliador. Quanto mais simples a tarefa, mais fácil será termos um entendimento generalizado dela e um maior acordo entre os juízes [Yang et al. 2016]. Em termos práticos, na Era das LLMs, deve-se **evitar pedir as justificativas (*rationales*)** dos juízes [Madsen 2024], especialmente quando essa justificativa não tem um objetivo claro dentro do ciclo de desenvolvimento.

A Teoria da Carga Cognitiva (*Cognitive Load Theory*) [Sweller 1988], usada em educação, design instrucional e usabilidade, incentiva que tarefas evitem informação irrelevante ou complexidade desnecessária, dada a capacidade limitada da memória de trabalho humana [Chen et al. 2018]. Se uma tarefa impõe uma carga cognitiva muito alta, o desempenho e a aprendizagem são prejudicados [de Jong 2010].

Tendemos a supor que pedir uma explicação lógica para uma resposta garante que a avaliação seja correta, mas não é bem assim. Tanto pessoas quanto modelos de linguagem costumam criar justificativas que parecem convincentes para a primeira resposta que possuem. Os raciocínios expostos geralmente não refletem o processo real que levou à resposta. Isso confunde plausibilidade (*plausibility*) com fidedignidade (*faithful-*

ness) [Jacovi and Goldberg 2020]. Além disso, avaliadores humanos tendem a ser enviesados por explicações geradas automaticamente e aceitá-las mesmo que haja outra justificativa mais apropriada para a tarefa. Pedir justificativas detalhadas também é caro e demorado e não garante mais precisão [Madsen 2024]. Por isso, esse tipo de explicação só deve ser solicitado quando houver clareza sobre seu uso e quando fizer parte do ciclo de desenvolvimento do sistema.

Talvez esse seja o Princípio mais controverso nos dias de hoje, com tantos trabalhos na área perseguindo *chain-of-thoughts, rationales* e *explanations*. Porém, a Psicologia Cognitiva é clara quanto aos processos mentais implícitos no julgamento humano e trabalhos recentes investigando profundamente o uso de explicações no contexto de LLMs também já demonstram a ineficiência desse tipo de argumento *post hoc*.

[Fayyaz et al. 2024] investigam alinhamento humano e fidelidade de *rationales* produzidos por LLMs utilizando diferentes tarefas. Os autores comparam explicações geradas com métodos de atribuição baseados em gradientes e atenção interna do modelo buscando relacionar os raciocínios textuais ao que acontece internamente nos modelos. Os experimentos utilizam datasets com *rationales* humanos anotados e avaliam LLMs robustas (Llama-2, Llama-3, Mistral, GPT-3.5 e GPT-4-Turbo). A avaliação de alinhamento humano é realizada por meio de F1 entre *rationales* humanos e *rationales* gerados, enquanto a fidelidade é medida via *flip rate* (frequência com que a remoção das palavras consideradas importantes altera a decisão do modelo). Os resultados mostram que *rationales* produzidos apresentam maior plausibilidade discursiva, mas menor fidelidade ao processo decisório interno do modelo quando comparados a métodos baseados em atribuição. Por exemplo, *rationales* gerados pelo Llama-3 apresentam um *flip rate* de apenas 22.33%. Esses resultados indicam que explicações produzidas por LLMs podem maximizar plausibilidade para avaliadores humanos, mas não refletem fatores responsáveis pela decisão do modelo. Assim, a exigência de *rationales* durante avaliações pode introduzir complexidade discursiva adicional sem garantir aumento correspondente de fidelidade ou confiabilidade. Além disso, oferecer aos avaliadores humanos explicações que não refletem os processos internos dos modelos pode induzi-los ao Viés de Acquiescência ou promover o Efeito de Ancoragem, já discutidos.

4.5. Princípio da Clareza

O Princípio da Clareza estabelece a necessidade de termos **diretrizes claras** sobre a avaliação a ser realizada pelo juiz. Em termos práticos, isso significa (i) ter diretrizes (ii) não ambíguas e (iii) com exemplos e casos de uso.

Prática já estabelecida na linguística de corpus e na anotação de dados em PLN [Freitas 2024], a necessidade de termos instruções claras para a avaliação das saídas das LLMs também já é reconhecida pela indústria[4].

O trabalho de [Pradhan et al. 2021] propõe uma construção iterativa para reduzir ambiguidades nas diretrizes de captura de julgamento, o *framework* FIND-Resolve-Label. Ele consiste em identificar dados ambíguos, resolver os casos e atualizar as diretrizes. A tarefa testada é simples, classificação binária de imagens, realizada com dezenas de anotadores do Mechanical Turk. Apesar da simplicidade da tarefa, a estratégia proposta

[4]`https://www.twine.net/blog/what-is-an-llm-evaluation-rubric/`
`https://www.ai21.com/glossary/foundational-llm/llm-as-a-judge/`

consegue aumentar a acurácia da anotação de aproximadamente 70% para 90%, medida contra um *ground truth* não disponível para os anotadores.

Quanto à captura de avaliações humanas, o trabalho de [Laux et al. 2024] apresenta um experimento com 307 anotadores variando as diretrizes recebidas (regras, regras incompletas e padrões vagos). O estudo também investiga o impacto do retorno financeiro da tarefa e revela que a acurácia dos julgamentos aumentou em 14% quando os humanos têm diretrizes claras em comparação a quando recebem apenas padrões. Em relação ao retorno financeiro, a diferença na acurácia é de apenas 2 pontos. Outro experimento em anotação de inferência em linguagem natural também revela que diretrizes detalhadas, com exemplos e definições técnicas dos fenômenos investigados, melhoram em 17% a anotação de relações semânticas como acarretamento e contradição [Kalouli et al. 2019].

Considerando LLMs, [Agarwal et al. 2024] investiga o papel dos exemplos nos prompts, explorando estratégias de *many-shots* que chegam a ter mais de mil exemplos, testados com o modelo Gemini 1.5 Pro, cuja janela de contexto chega a 1 milhão de tokens. Em todas as 11 tarefas investigadas, incluindo tradução automática e análise de sentimentos, a performance do modelo aumenta quando são oferecidos mais exemplos.

Em resumo, para qualquer tarefa que necessite da captura de julgamento, seja ele humano ou não, é necessário ter diretrizes e quanto mais claras elas forem, menos ruídos ou erros de julgamento estarão presentes nos dados finais.

5. Considerações e Conclusões

Visando criar uma ponte entre áreas que tradicionalmente estudam o conhecimento humano e a literatura sobre avaliação de sistemas baseados em PLN/IA, apresentamos cinco princípios simples, resumidos na Tabela 1. Estes princípios, baseados em conhecimento prévio e consistentes com outras áreas, devem ser integrados à prática de avaliação em PLN/IA. Na Tabela 1, encontram-se também possíveis formas de implementar estes princípios na prática, embora estas implementações recomendadas não sejam exaustivas.

Tabela 1. Resumo dos Princípios Metodológicos para Avaliação de Sistemas Baseados em IA

Princípio	Problema Mitigado	Implementação Recomendada
Independência	Viés de ancoragem, falso consenso, acquiescência e agreement bias	Avaliações às cegas
Diversidade	Viés individual ou cultural; inconsistências de julgamento	Uso de múltiplos avaliadores humanos e/ou múltiplas LLMs-juízes
Confiabilidade	Baixa reprodutibilidade e inconsistências	Mensuração contínua de acordo inter-anotadores (IAA)
Simplicidade	Sobrecarga cognitiva, racionalizações artificiais e baixa consistência	Tarefas objetivas e simples; evitar pedidos desnecessários de explicações
Clareza	Ambiguidade interpretativa e divergência de critérios	Criação de diretrizes explícitas, não ambíguas, com exemplos e casos limítrofes

Considerando que a avaliação em PLN/IA é fundamentalmente um problema de captura de julgamento humano, este trabalho focou em aspectos epistemológicos, metodológicos e operacionais no que toca à metodologia de captura do julgamento, seja este humano ou não. Oferecemos maneiras efetivas de operacionalizar a avaliação. Esperamos que este trabalho seja útil à comunidade, inspirando avaliações mais robustas e fidedignas.

6. Limitações

Neste trabalho, abordamos aspectos operacionais de captura de julgamento para avaliação. Esses aspectos são relevantes tanto em abordagens centradas no humano quanto em abordagens integralmente automatizadas. No entanto, aspectos muito relevantes da avaliação não foram abordados neste texto: taxonomia, esquemas, representatividade e *sampling*, ferramentas de gerenciamento ou versionamento, rastreabilidade, fluxo de trabalho de anotação (iterativo ou não), calibração de avaliadores, casos limítrofes e adversários, governança dos dados, documentação, detecção e mitigação de demais vieses, transparência, responsabilidade e reusabilidade. Também não abordamos questões fundamentais da avaliação humana como a definição do humano-referência (o especialista, o usuário, a população afetada pela aplicação) e as condições de trabalho dos anotadores, como remuneração justa e tempo adequado para o trabalho. Quanto à avaliação automática, aspectos como os impactos ambientais da etapa e a responsabilização pela qualidade da avaliação automática ficaram fora do escopo deste trabalho.

Este trabalho não inclui validação empírica dos princípios propostos, configurando-se antes como um artigo de posicionamento baseado em fundamentos da área de linguística de *corpus* e áreas correlatas que se dedicam à captura da percepção humana e à estruturação dessa percepção em grandes conjuntos de dados *machine-readable*. Embora muito dos princípios encontrem referencial teórico em diferentes áreas do conhecimento, nenhuma dessas áreas foi exaustivamente explorada; conceitos relevantes para as mesmas podem não ter sido investigados integralmente. Trabalhos futuros devem operacionalizar e mensurar o impacto da adoção destes princípios em pipelines reais de avaliação, especialmente na avaliação automatizada. Ressaltamos, porém, que validar cada um destes princípios não será trivial, exigindo experimentos extensos, abrangendo diferentes domínios e tarefas de avaliação, idealmente em diferentes línguas, com a participação de juízes distintos, com base em metodologia rigorosa de comparação.

7. Ética e Uso de Inteligência Artificial

Entendemos que esta pesquisa lança bases para uma metodologia integrada de avaliação de conteúdos gerados por IA, especialmente baseada em conhecimento anterior, provindo de áreas de expertise diferentes da IA em si. Neste contexto, é importante considerar que há inúmeras outras formas de pesquisa que não foram consideradas. Apesar de termos tentado ser abrangentes, é possível, e esperado, que ela não seja exaustiva. Há ainda muitos métodos e muitas áreas a explorar.

Durante esta pesquisa, utilizamos o assistente do Gemini para refinar buscas bibliográficas, o que se mostrou muito útil para conhecer princípios e nos aprofundar em áreas desconhecidas. Utilizamos também ferramentas de IA para validar nossas referências bibliográficas.

O texto deste artigo foi integralmente escrito por seres humanos.

8. Agradecimentos

Trabalho parcialmente apoiado pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) Código Financeiro 001; pela Fundação de Amparo à Pesquisa do Estado do Amazonas (FAPEAM) por meio do programa POSGRAD 2025, pelo Projeto NeuralBond (UNIVERSAL 2023, Proc. 01.02.016301.04300/2023-04); pelo CNPq no âmbito do INCT IAIA (406417/2022-9), INCT TILDIAR (408490/2024-1), Projeto AIDEn (Proc. 400936/2025-9) e por bolsa Produtividade em Pesquisa para Altigran da Silva (301925/2025-9).

Referências

Abercrombie, G., Dinkar, T., Curry, A. C., Rieser, V., and Hovy, D. (2025). Consistency is key: Disentangling label variation in natural language processing with intra-annotator agreement.

Agarwal, R., Singh, A., Zhang, L. M., Bohnet, B., Chan, S., Anand, A., Abbas, Z., Nova, A., Co-Reyes, J. D., Chu, E., Behbahani, F., Faust, A., and Larochelle, H. (2024). Many-shot in-context learning. *arXiv preprint arXiv:2404.11018*.

Artemova, E., Tsvigun, A., Schlechtweg, D., Fedorova, N., Tilga, S., and Obmoroshev, B. (2025). Hands-on tutorial: Labeling with llm and human-in-the-loop. In *Proceedings of COLING 2025*.

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Asli, C., Clark, E., and Gao, J. (2020). Evaluation of text generation: A survey. *arXiv:2006.14799*.

Bayerl, P. S. and Paul, K. I. (2011). What determines inter-coder agreement in manual annotations? a meta-analytic investigation. *Computational Linguistics*, 37(4):699–725.

Beloff, H. (1958). Two forms of social conformity: Acquiescence and conventionality. *The Journal of Abnormal and Social Psychology*, 56(1):99–104.

Bird, S. and Liberman, M. (2001). A formal framework for linguistic annotation. *Speech Communication*, 33(1–2):23–60.

Blagec, K., Dorffner, G., Moradi, M., Ott, S., and Samwald, M. (2022). A global analysis of metrics used for measuring performance in natural language processing. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pages 52–63, Dublin, Ireland. Association for Computational Linguistics.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Lisbon, Portugal. Association for Computational Linguistics.

BRASIL. Conselho Nacional de Justiça (2025). *Resolução n 615, de 10 de fevereiro de 2025*. Conselho Nacional de Justiça, Brasília, DF.

Callison-Burch, C., Osborne, M., and Koehn, P. (2006). Re-evaluating the role of bleu in machine translation research. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Chen, D., Chen, R., Zhang, S., Wang, Y., Liu, Y., Zhou, H., Zhang, Q., Wan, Y., Zhou, P., and Sun, L. (2024). MLLM-as-a-judge: Assessing multimodal LLM-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*.

Chen, O., Castro-Alonso, J., Paas, F., and Sweller, J. (2018). Extending cognitive load theory to incorporate working memory resource depletion: Evidence from the spacing effect. *Educational Psychology Review*, 30(2):483–501.

Choi, J., Hong, Y., and Kim, B. (2025). People will agree what i think: Investigating LLM's false consensus effect. In *Findings of the Association for Computational Linguistics: NAACL 2025*. Association for Computational Linguistics. arXiv:2407.12007v2.

Confident AI (2024). Deepeval: A framework for evaluating large language models. Accessed: 2026-05-12.

de Jong, T. (2010). Cognitive load theory, educational research, and instructional design: Some food for thought. *Instructional Science*, 38(2):105–134.

Durkin, K. (1996). Peer pressure. In Manstead, A. S. R. and Hewstone, M., editors, *The Blackwell Encyclopedia of Social Psychology*. Blackwell Publishers, Oxford.

Fanous, A., Goldberg, J., Agarwal, A. A., Lin, J., Zhou, A., Daneshjou, R., and Koyejo, S. (2025). Syceval: Evaluating llm sycophancy. *arXiv preprint arXiv:2502.08177*.

Fayyaz, M., Yin, F., Sun, J., and Peng, N. (2024). Evaluating human alignment and model faithfulness of llm rationale.

Foody, G. M. (2024). Ground truth in classification accuracy assessment: Myth and reality. *Geomatics*, 4(1):81–90.

Fort, K., Adda, G., and Cohen, K. B. (2011). Last words: Amazon mechanical turk: Gold mine or coal mine? *Computational Linguistics*, 37(2):413–420.

Freitas, C. (2024). Dataset e corpus. In *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, chapter 16. BPLN, 4 edition.

Frenda, S., Abercrombie, G., Basile, V., Pedrani, A., Panizzon, R., Cignarella, A. T., Marco, C., and Bernardi, D. (2024). Perspectivist approaches to natural language processing: a survey: Perspectivist approaches to natural language processing... *Lang. Resour. Eval.*, 59(2):1719–1746.

Gao, M., Hu, X., Yin, X., Ruan, J., Pu, X., and Wan, X. (2025). LLM-based NLG evaluation: Current status and challenges. *Computational Linguistics*, 51:661–687.

Gilbert, D. T. (1991). How mental systems believe. *American Psychologist*, 46(2):107–119.

Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., and Kamar, E. (2022). Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection.

Jacovi, A. and Goldberg, Y. (2020). Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4212, Online. Association for Computational Linguistics.

Jain, S., Ahmed, U. Z., Sahai, S., and Leong, B. (2025). Beyond consensus: Mitigating the agreeableness bias in llm judge evaluations.

Jang, M. E. and Silavong, F. (2025). InstaJudge: Aligning judgment bias of LLM-as-judge with humans in industry applications. In Potdar, S., Rojas-Barahona, L., and Montella, S., editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1158–1172, Suzhou (China). Association for Computational Linguistics.

Kalouli, A.-L., Buis, A., Real, L., Palmer, M., and de Paiva, V. (2019). Explaining simple natural language inference. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 132–143, Florence, Italy. Association for Computational Linguistics.

Kalouli, A.-L., Real, L., and de Paiva, V. (2017). Textual inference: getting logic from humans. In Gardent, C. and Retoré, C., editors, *Proceedings of the 12th International Conference on Computational Semantics (IWCS) — Short Papers*, Montpellier, France. Association for Computational Linguistics.

Laux, J., Stephany, F., and Liefgreen, A. (2024). Improving task instructions for data annotators: How clear rules and higher pay increase performance in data annotation in the ai economy.

Lee, Y., Kim, J., Kim, J., Cho, H., Kang, J., Kang, P., and Kim, N. (2025). Checkeval: A reliable llm-as-a-judge framework for evaluating text generation using checklists. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Suzhou, China. Association for Computational Linguistics.

Li, K., Li, Y., Zhang, T., Luo, H., Wu, X., Glass, J. R., and Meng, H. M. (2025a). Ragzeval: Enhancing rag response evaluation through end-to-end reasoning and ranking-based reinforcement learning. In *Proceedings of EMNLP 2025*.

Li, S., Li, J., Qu, Y., Shi, X., Guo, Y., He, Z., Wang, Y., and Tan, W. (2025b). Semantic-eval : A semantic comprehension evaluation framework for large language models generation without training. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T., editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9675–9690, Vienna, Austria. Association for Computational Linguistics.

Lin, S., Hilton, J., and Evans, O. (2021). Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Madsen, A. (2024). New faithfulness-centric interpretability paradigms for natural language processing. *arXiv preprint arXiv:2411.17992*.

Marasović, A. (2018). Nlp's generalization problem, and how researchers are tackling it. *The Gradient*. Accessed: 2026-05-14.

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., and Zamparelli, R. (2014). A SICK cure for the evaluation of compositional distributional semantic mo-

dels. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Marinho, M., Vianna, D., Migliorini, G., Real, L., Ramalhete, M., and da Silva, A. (2026). Grounding the evaluation of ai-generated legal answers: The hermeval protocol and empirical evidence. In *Proceedings of the International Conference on Artificial Intelligence and Law (ICAIL)*, Singapore. ACM.

Martin, D., Hanrahan, B. V., O'Neill, J., and Gupta, N. (2014). Being a turker. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, page 224–235, New York, NY, USA. Association for Computing Machinery.

McEnery, T. and Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press, Cambridge, 2 edition.

Messick, S. and Jackson, D. N. (1961). Acquiescence and the factorial interpretation of personality tests. In Jackson, D. N. and Messick, S., editors, *Psychological Measurement in Personality Research*, pages 61–103. University of Chicago Press, Chicago.

Meyer, C. F. (2002). *English Corpus Linguistics: An Introduction*. Cambridge University Press, Cambridge.

Novikova, J., Dušek, O., Cercas Curry, A., and Rieser, V. (2017). Why we need new evaluation metrics for NLG. In Palmer, M., Hwa, R., and Riedel, S., editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Pagano, T. P., Loureiro, R. B., Lisboa, F. V. N., Peixoto, R. M., Guimarães, G. A. S., Cruz, G. O. R., Araujo, M. M., Santos, L. L., Cruz, M. A. S., Oliveira, E. L. S., Winkler, I., and Nascimento, E. G. S. (2023). Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big Data and Cognitive Computing*, 7(1).

Pavlick, E. and Kwiatkowski, T. (2019). Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Plank, B. (2022). The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022)*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pradhan, V. K., Schaekermann, M., and Lease, M. (2021). In search of ambiguity: A three-stage workflow design to clarify annotation guidelines for crowd workers.

Qian, M., Sun, G., Gales, M. J. F., and Knill, K. M. (2026). Who can we trust? llm-as-a-jury for comparative assessment.

Raina, V., Liusie, A., and Gales, M. (2024). Is llm-as-a-judge robust? investigating universal adversarial attacks on zero-shot llm assessment. *arXiv preprint arXiv:2402.14016*.

Reiter, E. and Belz, A. (2009). An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

Ross, L., Greene, D., and House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, 13(3):279–301.

Sandan, I. B., Dinh, T. A., and Niehues, J. (2025). Knockout LLM assessment: Using large language models for evaluations through iterative pairwise comparisons. In Arviv, O., Clinciu, M., Dhole, K., Dror, R., Gehrmann, S., Habba, E., Itzhak, I., Mille, S., Perlitz, Y., Santus, E., Sedoc, J., Shmueli Scheuer, M., Stanovsky, G., and Taffjord, O., editors, *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 121–128, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fast—but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP 2008*, Honolulu, Hawaii. Association for Computational Linguistics.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285.

Thakur, A. S., Choudhary, K., Ramayapally, V. S., Vaidyanathan, S., and Hupkes, D. (2024). Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. *arXiv preprint arXiv:2406.12624*.

Trautmann, D., Ostapuk, N., Grail, Q., Pol, A. A., Bonifazi, G., Gao, S., and Gajek, M. (2024). Measuring the groundedness of legal question-answering systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.

Verga, P., Hofstätter, S., Althammer, S., Su, Y., Piktus, A., Arkhangorodsky, A., Xu, M., White, N., and Lewis, P. (2024). Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*.

Voutilainen, A. (2012). Improving corpus annotation productivity: a method and experiment with interactive tagging. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2097–2102, Istanbul, Turkey. European Language Resources Association (ELRA).

Wei, H., He, S., Xia, T., Liu, F., Wong, A., Lin, J., and Han, M. (2025). Systematic evaluation of llm-as-a-judge in llm alignment tasks: Explainable metrics and diverse prompt templates.

Wojcieszak, M. and Price, V. (2009). What underlies the false consensus effect? how personal opinion and disagreement affect perception of public opinion. *International Journal of Public Opinion Research*, 21(1):25–46.

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., and He, L. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381.

Yang, J., Redi, J., Demartini, G., and Bozzon, A. (2016). Modeling task complexity in crowdsourcing. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 4(1):249–258.

Zhao, W. X., Zhou, K., Li, J., Tang, T., and Wen, J.-R. (2026). *Human Alignment*, pages 209–256. Springer Nature Singapore, Singapore.

Zheng, L., Chiang, W.-L., Sheng, Y., and et al. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS Workshop on Instruction Tuning and Instruction Following*.

Zhu, Z., Tieleman, O., Bukhtiyarov, A., and Chen, J. (2026). Cyclicjudge: Mitigating judge bias efficiently in llm-based evaluation.

Clonagem de Voz por IA para o Português Brasileiro em Contexto Assistivo: Desafios Comparativos em Cinco Ferramentas de Síntese de Fala

Luiz G. dos Santos[1], Nathalia Borges[2], Matheus Sales[2], Luciana C. L. F. Borges[1], Thais R. Kempner[2], Eunice P. S. Nunes[1]

[1]Instituto de Computação – Universidade Federal de Mato Grosso (UFMT), Cuiabá – MT – Brazil; [2]Faculdade de Engenharia – Universidade Federal de Mato Grosso (UFMT), Cuiabá – MT – Brazil

luizgdsantos91@gmail.com, nathalia@sou.ufmt.br, matheus2017mbs@gmail.com, lucianafariaborges@gmail.com, thaisrgk@gmail.com, eunice@ufmt.br

Abstract. Neural TTS models underperform for Brazilian Portuguese (PT-BR) regarding child voices, emotional prosody, and sublexical phonemes. This paper evaluates five AI voice synthesis tools in an assistive context using 170 studio recordings as corpus. Two challenges persisted across all approaches: emotional prosody and PT-BR phonemes absent from English. Incorporating the multilingual-phonemes-10k-alpha dataset into StyleTTS 2 consistently improved regional phonetic quality. Findings provide a comparative assessment and evidence of a low-cost strategy to improve PT-BR phonetic coverage.

Resumo. Modelos de síntese neural de fala apresentam desempenho reduzido para o português brasileiro (PT-BR) em vozes infantis, prosódia emocional e fonemas sublexicais. Este trabalho avalia cinco ferramentas de síntese por IA em contexto assistivo, com 170 gravações como corpus controlado. Dois desafios persistiram: prosódia emocional e fonemas do PT-BR ausentes do inglês. A incorporação do dataset multilingual-phonemes-10k-alpha ao StyleTTS 2 produziu melhoria consistente na qualidade fonética regional. Os achados oferecem avaliação comparativa e evidência de estratégia de baixo custo para ampliar a cobertura fonética do PT-BR.

1. Introdução

Apesar dos avanços recentes na síntese neural de fala, o desempenho observado em inglês não se transfere adequadamente para línguas de médio recurso como o português brasileiro (PT-BR), criando um descompasso entre estado da arte e aplicações reais. Essa assimetria decorre da escassez de datasets públicos em PT-BR com diversidade de falantes e anotação prosódica adequada [Casanova et al. 2022], e é agravada pelo fato de que modelos treinados predominantemente em inglês tendem a degradar a qualidade para vozes fora dos perfis de referência [Araújo et al. 2026]. Em aplicações assistivas, a autenticidade vocal impacta diretamente a efetividade terapêutica, pois variações de

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 348–360
October 19-22, 2026 ©2026 Association for Computational Linguistics

timbre e entonação afetam o engajamento e a compreensão emocional em crianças com perfis neurodivergentes [Otto-Meyer et al. 2018].

A literatura ainda não oferece avaliações sistemáticas que integrem a síntese de voz infantil, a adaptação ao PT-BR em cenários de dados limitados e a reprodução fiel de características fonéticas e prosódicas em aplicações reais. Nesse contexto, este trabalho investiga essa lacuna a partir de um caso concreto: a necessidade de clonar e expandir o banco de voz de um robô educacional para tutoria infantil com crianças com Transtorno do Espectro Autista (TEA), Transtorno do Déficit de Atenção com Hiperatividade (TDAH) e Dislexia, população para a qual robôs sociais têm eficácia documentada como mediadores terapêuticos [Scassellati et al. 2012]. O banco original foi gravado em estúdio por um menino de 11 anos, totalizando 170 enunciados de vocabulário cotidiano, expressões emocionais e conteúdo educacional [Rebouças et al. 2021]. Com a mudança vocal decorrente da puberdade, tornou-se inviável manter a consistência tímbrica em novas gravações, o que motivou a investigação de clonagem de voz por IA. O sucessor pedagógico do robô, o BITI (Brinquedo Inclusivo para Treinamento Infantil), voltado para alfabetização pelo método fônico [Oliveira e Albuquerque 2021], acrescenta o requisito de fonemas isolados, uma classe de entrada para a qual os modelos neurais de TTS não foram originalmente projetados.

Diante dessas restrições, este trabalho tem como objetivo avaliar a capacidade de diferentes abordagens de síntese e clonagem de voz em reproduzir uma voz infantil em PT-BR com naturalidade e fidelidade fonética, considerando simultaneamente três dimensões críticas: clonagem tímbrica, controle prosódico emocional e síntese de fonemas sublexicais. Para isso, foi conduzido um estudo empírico comparativo ao longo de 18 meses, envolvendo cinco ferramentas.

As principais contribuições deste trabalho são: (i) uma avaliação comparativa controlada de cinco abordagens de síntese de fala aplicadas à voz infantil em PT-BR em um contexto assistivo real; (ii) a análise sistemática dos desafios fonológicos específicos do PT-BR em modelos majoritariamente treinados em inglês; e (iii) a demonstração do impacto da incorporação do dataset multilingual-phonemes-10k-alpha na melhoria da qualidade fonética regional, um resultado ainda pouco explorado na literatura. Os achados têm relevância prática direta para grupos que trabalham com síntese de fala em PT-BR sob restrição de dados, oferecendo tanto um mapa comparativo de ferramentas disponíveis quanto uma estratégia de baixo custo para ampliar a cobertura fonética regional.

2. Metodologia

A partir do objetivo de avaliar o desempenho de cinco ferramentas de síntese e clonagem de voz na reprodução de fala infantil em PT-BR sob severa restrição de dados, este trabalho, que se configura em um estudo experimental comparativo de natureza aplicada, foi conduzido ao longo de 18 meses. O estudo foi estruturado para analisar sistematicamente a fidelidade tímbrica ao locutor original, o controle prosódico

emocional e a precisão fonêmica, com ênfase em segmentos característicos do PT-BR. Todas as ferramentas foram submetidas às mesmas condições de entrada, incluindo corpus de treinamento, textos de teste e critérios de avaliação, de modo a permitir comparação equivalente. Os experimentos foram conduzidos em máquina com processador Ryzen 5 5500, GPU AMD RX 6750 XT (12 GB VRAM), 16 GB de RAM e armazenamento NVMe de 1 TB, executando Ubuntu 24.04.

2.1. Fundamentação na Literatura

O TTS-Portuguese Corpus [Casanova et al. 2022], com 10 horas e 28 minutos de um único falante adulto, foi por anos um dos principais recursos públicos para síntese end-to-end em PT-BR, estabelecendo MOS de referência de 4,03 com Tacotron 2. O panorama de recursos públicos de fala em PT-BR é hoje mais amplo, incluindo corpora multilocutor como o Multilingual LibriSpeech e o Common Voice, além de recursos mais recentes como o Certas Palavras [Araújo et al. 2026] e o capítulo de recursos de fala do livro PLN-BR [Casanova et al. 2026]; ainda assim, poucos desses recursos combinam falante único, qualidade de estúdio e pontuação textual — condições relevantes para síntese de fala. Entretanto, o desempenho originalmente reportado para o TTS-Portuguese Corpus não se sustenta em condições mais próximas do uso real [Araújo et al. 2026]. O modelo multilíngue YourTTS foi treinado com apenas um falante em português [Casanova et al. 2024], implicando degradação para vozes atípicas como a infantil com variação regional.

No campo da prosódia, Barakat et al. (2024), em revisão sistemática com 356 artigos, mostram que modelos de transferência de estilo suavizam extremos emocionais. Bott et al. (2024) propõem prompts em linguagem natural como alternativa de controle. Para fonemas sublexicais, Jurafsky e Martin (2023) explicam que modelos de TTS assumem coarticulação sentencial, tornando fonemas isolados uma entrada fora do regime de treinamento. Maia et al. (2023) confirmam que a geração prosódica é o componente mais frágil em TTS para PT-BR. Em conjunto, esses trabalhos evidenciam limitações recorrentes que a literatura ainda não integrou em avaliações sistemáticas sob restrição de dados, lacuna que orienta este estudo.

2.2. Corpus e Preparação dos Dados

O corpus base é composto por 170 enunciados gravados em estúdio por um locutor masculino de 11 anos com leve sotaque carioca, totalizando aproximadamente 2 minutos e 50 segundos de áudio [Rebouças et al. 2021]. O conjunto combina enunciados de palavra isolada (por exemplo, “alface”, “amarelo”, palavras-veículo como “b de bola”) com frases curtas de vocabulário cotidiano, expressões emocionais e conteúdo educacional, o que explica a curta duração média por enunciado. As gravações apresentam alta qualidade acústica; o áudio foi segmentado manualmente por enunciado, normalizado em intensidade e acompanhado de transcrições textuais revisadas. Quando necessário, os arquivos foram replicados em loop até atingir o volume mínimo exigido

pelas ferramentas, conforme recomendado na documentação do Piper [OHF-Voice 2025]. Em etapas posteriores, amostras sintéticas aprovadas pela equipe foram incorporadas para ampliar a cobertura fonética. Para experimentos de melhoria fonética em PT-BR, foi incorporado ao pipeline o dataset multilingual-phonemes-10k-alpha [StyleTTS2 Community 2024], com ênfase em segmentos sub-representados nos modelos avaliados.

2.3. Ferramentas e Paradigmas Avaliados

Foram avaliadas cinco ferramentas representativas de paradigmas distintos. O ElevenLabs [ElevenLabs 2024] é uma API comercial baseada em nuvem. O Coqui TTS com XTTS-v2 [Casanova et al. 2024] é um modelo open source de clonagem zero-shot multilíngue. O Piper TTS [OHF-Voice 2025] é um modelo VITS com fine-tuning supervisionado. O StyleTTS 2 [Li et al. 2023] e o Kokoro-82M [Hexgrad 2024] são modelos baseados em difusão de estilo. Por fim, o F5-TTS [Chen et al. 2024] combinado com RVC [RVC Project 2023] constitui uma abordagem híbrida de clonagem few-shot com conversão de voz. Cabe notar que StyleTTS 2 e Kokoro-82M são tratados neste trabalho como um único paradigma (difusão de estilo), mas reportados separadamente na Tabela 1 e na Seção 3.4 por apresentarem tempos de convergência e comportamentos distintos o suficiente para justificar análise individualizada. Todas foram configuradas para utilizar o mesmo corpus base, com ajustes mínimos para viabilizar a convergência em ambiente computacional disponível.

2.4. Protocolo de Treinamento

Os procedimentos de treinamento variaram conforme o paradigma de cada ferramenta e foram escolhidos com base na melhor adequação ao corpus disponível e às restrições do projeto. Para clonagem zero-shot (XTTS-v2, F5-TTS), as amostras de referência foram utilizadas diretamente, escolha justificada pela necessidade de clonagem rápida com volume reduzido de dados. Para fine-tuning supervisionado (Piper, StyleTTS 2, Kokoro-82M), os modelos foram adaptados ao corpus no formato requerido, abordagem adotada por permitir maior especialização ao timbre e padrão prosódico do locutor-alvo. Para o ElevenLabs, foi criado um perfil de voz via interface, dado que a plataforma não permite acesso aos pesos do modelo.

Nos modelos com treinamento iterativo, o processo foi monitorado por inspeção qualitativa periódica das amostras geradas. O treinamento iterativo foi conduzido até a estabilização da qualidade da síntese, definida como o ponto a partir do qual melhorias adicionais não eram detectáveis de forma consistente. Empiricamente, esse ponto ocorreu em torno de 15 horas para a maioria das arquiteturas. O tempo mínimo até resultado utilizável variou consideravelmente: RVC convergiu em 10 a 45 minutos; StyleTTS 2 e Kokoro-82M entre 1 e 3 horas; Piper TTS e F5-TTS entre 2 e 5 horas; e o XTTS-v2 demandou entre 24 e 72 horas para o mesmo corpus — tempo mais elevado que se explica pela escassez de dados em português no treinamento original do modelo, o que exige

maior adaptação da distribuição aprendida à nova voz (discutido em mais detalhe na Seção 5).

2.5. Protocolo de Avaliação Qualitativa

A avaliação dos áudios sintetizados foi conduzida por um painel de quatro avaliadores pertencentes à equipe de pesquisa, todos com experiência prévia em processamento de fala e interação humano-robô. Cada amostra foi avaliada segundo quatro critérios qualitativos. O primeiro foi a naturalidade, entendida como grau de semelhança com a fala humana do locutor original. O segundo foi a inteligibilidade, ou seja, a clareza e a compreensão do conteúdo pelo público-alvo. O terceiro foi a coerência emocional, referente à adequação da entonação ao estado afetivo pretendido. O quarto foi a precisão fonêmica, que avalia a fidelidade na realização de fonemas do PT-BR, com atenção a segmentos críticos. A aprovação exigia consenso entre os avaliadores em todos os critérios simultaneamente.

2.6. Protocolo para Fonemas Sublexicais

Para a análise de fonemas isolados exigidos pelo método fônico do robô educacional, foi definido um conjunto de segmentos-alvo representativos, incluindo consoantes plosivas (/b/, /d/, /p/, /t/), fricativas (/f/, /v/, /s/, /z/) e líquidas (/l/, /lh/, /nh/, /R/). A seleção foi orientada pelos fonemas de aquisição mais tardia na criança, que concentram as maiores dificuldades de aprendizado de leitura em crianças com dislexia e TEA, e pelos segmentos com maior probabilidade de falha nos modelos avaliados, especialmente aqueles sem equivalente no inglês. Como os modelos de TTS assumem contexto sentencial, foram testadas duas abordagens: a entrada direta do fonema como texto de síntese, para avaliar o comportamento do modelo diante de entrada sublexical, e a estratégia de palavra-veículo, com inserção do fonema em palavra de contexto controlado e extração manual do segmento em software de edição de áudio. A eficácia foi avaliada quanto à ausência de artefatos como inserção de vogais intrusivas, distorções ou substituições fonêmicas.

3. Ferramentas Avaliadas e Resultados Obtidos

3.1. ElevenLabs

O ElevenLabs é uma plataforma comercial de síntese e clonagem de voz em nuvem, avaliada por sua acessibilidade e suporte declarado ao português. Os controles expostos ao usuário limitam-se a três parâmetros (estabilidade, similaridade com a referência e exagero de estilo), cujo comportamento é não-linear. A geração é estocástica sem semente configurável, de modo que o mesmo texto com as mesmas configurações produz saídas distintas a cada execução.

Para fala neutra, os resultados foram satisfatórios. Para fala emocional, a ausência de controle prosódico direto obrigou o uso de estratégias tipográficas (reticências para ritmo mais lento, exclamações para animação, prefixos como “Com voz muito assustada:”),

com efeitos detectáveis em tristeza e cansaço, porém nulos para medo e surpresa. A calibração acumulada durante uma sessão não persistia entre solicitações de emoções distintas.

Para o método fônico, o ElevenLabs é inadequado. A plataforma não foi projetada para entrada sublexical e apresentou sistematicamente inserção de vogal intrusiva ou silêncio ao receber fonemas isolados como texto. A ausência de controle de semente aleatória torna ainda mais problemática a tarefa: mesmo que um fonema específico seja sintetizado corretamente em uma tentativa, não há garantia de reprodução nas seguintes.

3.2. Coqui TTS / XTTS-v2

O XTTS-v2 é um modelo de clonagem zero-shot multilíngue de código aberto que aceita referências de seis segundos e suporta português brasileiro [Casanova et al. 2024]. A instalação revelou-se mais complexa do que o esperado; incompatibilidades de versão entre PyTorch, torchaudio e CUDA demandaram dias de diagnóstico.

Para fala neutra, a qualidade foi satisfatória e a fidelidade tímbrica ao locutor de referência foi perceptível. Para fala emocional, o mecanismo de condicionamento por clip de referência, que consiste em fornecer um áudio original com a emoção desejada como âncora, atenuava sistematicamente os extremos prosódicos: “Estou com medo” entregue com voz tensa no clip chegava sintetizado próximo ao neutro. O problema decorre da ausência de separação explícita entre identidade do falante e estado emocional no espaço latente, conforme documentado por Cho et al. (2025): o modelo combina os dois sinais em vez de transferir apenas o componente emocional.

Para fonemas isolados, o modelo produziu os mesmos artefatos observados nas demais ferramentas: inserção de vogal intrusiva e troca de idioma no tokenizador multilíngue. A Coqui.ai encerrou as operações em dezembro de 2023; repositório é mantido por fork comunitário, o que representa riscos para projetos com horizonte longo.

3.3. Piper TTS

O Piper é uma ferramenta open source de síntese neural baseada em VITS, projetada para fine-tuning com corpus de falante único e exportação para ONNX Runtime [OHF-Voice 2025]. Diferentemente das ferramentas anteriores, não realiza clonagem zero-shot; requer treinamento supervisionado no formato LJSpeech. Foi avaliado pela estabilidade do pipeline de fine-tuning e pela execução completamente local.

Para fala neutra, o Piper produziu os resultados mais estáveis e reprodutíveis entre todas as ferramentas avaliadas: a mesma entrada de texto gerava saídas auditivamente consistentes entre execuções, o que facilita o controle de qualidade. O treinamento convergiu de forma mais estável do que o XTTS-v2 em situação equivalente, atingindo o platô de qualidade em torno de 15 horas.

Para fala emocional, a ausência de qualquer mecanismo de transferência de estilo ou condicionamento emocional implica síntese neutra por padrão, independentemente da formulação do texto de entrada.

Para o método fônico, o Piper é a ferramenta mais promissora entre as avaliadas, com ressalvas. A reprodutibilidade da saída e a ausência de dependência de GPU para inferência são vantagens diretas para a produção de um banco fonético estável. Contudo, fonemas isolados continuam sendo problemáticos na entrada direta, e a estratégia de palavra-veículo é mais eficaz com o Piper do que com os demais modelos justamente pela consistência entre gerações: o mesmo recorte pode ser reutilizado sem variação. A principal limitação para o BITI é a necessidade de retraining completo para incorporar novos fonemas, sem mecanismo de clonagem rápida.

3.4. StyleTTS 2 e Kokoro-82M

StyleTTS 2 [Li et al. 2023] e Kokoro-82M [Hexgrad 2024] são arquiteturas de difusão de estilo: em vez de condicionar sobre um clip de referência, modelam o estilo como variável latente amostrada durante a síntese, usando o WavLM como discriminador. Ambos foram treinados quase exclusivamente em inglês. O Kokoro, com 82 milhões de parâmetros, apresentou a convergência de fine-tuning mais rápida entre todos os modelos avaliados, atingindo o platô de qualidade em tempo inferior a 15 horas.

Para fala neutra em inglês, a qualidade de ambos é notável. Para fala em PT-BR, fonemas sem equivalente inglês não estavam adequadamente representados no espaço aprendido: vogais nasais (/ã/, /ẽ/, /ĩ/, /õ/, /ũ/) soavam mitigadas ou substituídas por aproximantes ingleses, e o ditongo /ão/ saía consistentemente comprometido. A incorporação do dataset multilingual-phonemes-10k-alpha [StyleTTS2 Community 2024] ao pipeline de treinamento do StyleTTS 2 produziu melhora perceptível e consistente na autenticidade fonética regional, descrita pelos avaliadores como audível. Esse resultado emergiu de uma modificação exploratória do pipeline e é discutido na Seção 5.

Para fala emocional, a difusão de estilo introduz variabilidade prosódica natural maior do que os modelos de referência. Contudo, sem controle explícito da dimensão afetiva no espaço latente, o estado emocional sintetizado é não determinístico e não corresponde necessariamente ao pretendido.

Para o método fônico, o StyleTTS 2 e o Kokoro apresentam o maior potencial futuro, mas a situação atual é limitada. Os fonemas nasais do PT-BR e o ditongo /ão/ são exatamente os segmentos mais relevantes para o método fônico brasileiro e os mais problemáticos nesses modelos. A incorporação do dataset multilíngue atenuou esse problema, mas não o eliminou. Ao mesmo tempo, a difusão de estilo permite maior variabilidade prosódica que pode ser explorada para a síntese de fonemas com entonação contextualmente adequada. Se o suporte ao PT-BR for ampliado nas versões futuras desses modelos, eles se tornarão os candidatos mais fortes para a aplicação fonética, dado o menor tempo de treinamento.

3.5. F5-TTS e Abordagem Híbrida com RVC

O F5-TTS [Chen et al. 2024] é um modelo de clonagem few-shot que opera por fluxo retificado sobre mel-espectrograma, requerendo apenas alguns segundos de referência. Para fala neutra, a fidelidade tímbrica foi satisfatória, comparável ao XTTS-v2. Para fala emocional, o problema de atenuação prosódica reproduziu-se com o mesmo padrão das ferramentas anteriores baseadas em referência. Para fonemas isolados, o comportamento foi igualmente problemático.

Em paralelo, foi investigada uma estratégia híbrida utilizando o RVC [RVC Project 2023], um conversor de identidade vocal, sobre saídas do XTTS-v2. O RVC recebe qualquer áudio de entrada e converte o timbre para a voz treinada, com alta fidelidade quando a entrada é voz humana gravada ao vivo. A estratégia proposta era usar o TTS para gerar o conteúdo textual e o RVC para transferir a identidade tímbrica. Na prática, quando a entrada para o RVC era saída de TTS, a artificialidade da síntese era amplificada pelo conversor, não atenuada. Com voz humana como entrada, os resultados melhoraram, mas a configuração vincula o pipeline ao timbre de uma pessoa específica, comprometendo a reprodutibilidade.

Para o método fônico, a abordagem RVC apresenta um problema específico: a neutralidade articulatória necessária para fonemas isolados depende de que o locutor de entrada produza o fonema de forma precisa e descontextualizada. Qualquer variação de entonação ou coarticulação na voz de entrada é transferida para a saída pelo conversor, tornando a padronização fonética difícil. A abordagem é ainda menos indicada para o BITI do que para o OTTO, pois os requisitos de precisão fonêmica são mais exigentes.

4. Desafios Transversais

4.1. Síntese Emocional

Os enunciados emocionais do banco original apresentam perfis prosódicos clinicamente distintos. “Estou com sono” inclui a qualidade vocal característica de bocejo, produzida pelo abaixamento do véu palatino, ampla abertura faríngea e ressonância nasal aumentada, resultando em timbre simultaneamente nasalizado e distendido. Nenhuma das cinco ferramentas reproduziu essa característica de forma reconhecível. “Estou com medo” exige tensão vocal e aceleração de ritmo que os modelos de transferência de estilo atenuavam sistematicamente. Trata-se de uma limitação com implicação terapêutica direta: o robô terapêutico é utilizado para treino de reconhecimento emocional com crianças que apresentam dificuldade nessa área [Otto-Meyer et al. 2018], e um áudio com expressão emocional fraca ou incorreta pode ser pedagogicamente inerte ou, em casos de hipersensibilidade sensorial, perturbador.

O problema é estrutural. Modelos treinados sem dados emocionais anotados para PT-BR sintetizam prosódia próxima ao centro da distribuição de entonação por padrão, suficiente

para fala neutra mas insuficiente para os extremos necessários à expressão emocional convincente, conforme previsto pela literatura [Barakat et al. 2024].

4.2. Fonemas Sublexicais e Nasalização

O método fônico exige áudios de fonemas isolados (/b/, /d/, /f/, entre outros) para exercícios de associação grafema-fonema no BITI. Essa entrada viola as premissas dos modelos neurais de TTS: Jurafsky e Martin (2023) explicam que esses modelos assumem contexto sentencial e coarticulação entre segmentos adjacentes, de modo que um fonema consonantal isolado produz resultados imprevisíveis, incluindo inserção de vogal intrusiva, silêncio ou troca de idioma no tokenizador multilíngue.

A estratégia parcialmente eficaz desenvolvida consistiu em embutir o fonema-alvo em uma palavra-veículo de contexto controlado e extrair o segmento via edição em software de áudio (por exemplo, sintetizar “afta” para isolar /f/, ou “bala” para /b/ inicial). A técnica é aplicável a fricativas intervocálicas, mas falha para /R/ vibrante múltipla, /lh/ e /nh/, três fonemas que continuam dependendo de gravação humana suplementar. Nos modelos baseados em inglês (StyleTTS 2, Kokoro), as vogais nasais e o ditongo /ão/ acrescentaram uma segunda camada de dificuldade, parcialmente mitigada pelo dataset multilingual-phonemes-10k-alpha [StyleTTS2 Community 2024].

A Tabela 1 sintetiza os resultados comparativos das cinco ferramentas nas dimensões avaliadas.

Ferramenta	Execução	Emoção	Fon. PT-BR	Método Fônico	Treino (platô)
ElevenLabs	API nuvem (pago)	Prompt (fraca)	Médio	Inadequado	N/A
XTTS-v2	Local, GPU ≥8GB	Ref. (atenuada)	Médio	Limitado**	~15h
Piper TTS	Local, CPU/GPU	Neutro	Médio	Mais indicado**	~15h
StyleTTS2+ML*	Local, GPU ≥8GB	Difusão latente	Bom*	Potencial futuro	~15h
Kokoro-82M	Local, GPU ≥8GB	Difusão latente	Médio	Potencial futuro	<15h
F5-TTS+RVC	Local, GPU ≥8GB	Ref. (atenuada)	Médio	Não indicado	~15h

*Tabela 1. Comparação das ferramentas avaliadas. N/A = dado não sistematizado nesta fase. * Com dataset multilingual-phonemes-10k-alpha. ** Estratégia de palavra-veículo parcialmente eficaz. *** StyleTTS2+ML e Kokoro-82M correspondem a um único paradigma (difusão de estilo), reportado em linhas separadas por apresentarem resultados distintos.*

5. Discussão

Os resultados deste estudo têm implicações que ultrapassam o contexto específico do BITI. A dificuldade em sintetizar prosódia emocional e fonemas nasais em PT-BR não é uma peculiaridade do caso avaliado: é uma consequência previsível de modelos treinados predominantemente em inglês, sem anotação emocional e sem representação adequada dos fonemas do português. Pesquisadores que trabalham com qualquer aplicação de

síntese de fala em PT-BR com requisitos além da neutralidade prosódica enfrentarão o mesmo conjunto de limitações, independentemente da ferramenta escolhida.

A assimetria nos tempos de treinamento revela algo importante sobre o custo real de adoção dessas ferramentas. O XTTS-v2, que exigiu entre 24 e 72 horas, corresponde a um modelo com poucos dados em português no treinamento original, o que implica que mais tempo de adaptação é necessário para ajustar a distribuição aprendida a uma nova voz. StyleTTS 2 e Kokoro, com 1 a 3 horas, chegaram ao platô mais rapidamente não porque são melhores, mas porque a difusão de estilo tem uma dinâmica de adaptação diferente. O platô em 15 horas, comum a todos os modelos com treinamento iterativo, indica que o corpus de 170 enunciados esgota sua capacidade de generalização em poucas horas. Essa conclusão tem implicação direta: ampliar o corpus é mais produtivo do que estender o treinamento.

O efeito do dataset multilingual-phonemes-10k-alpha é o resultado mais acionável deste trabalho para a comunidade. Trata-se de um recurso público, de incorporação simples ao pipeline de fine-tuning do StyleTTS 2, que produziu melhoria audível sem re-treinamento completo. A hipótese mais plausível é que os fonemas nasais do português presentes no dataset ampliam o espaço fonético coberto pelo modelo, reduzindo a tendência de substituição por aproximantes ingleses. Esse mecanismo, se confirmado por ablação, teria implicação para qualquer grupo que trabalhe com síntese em línguas cujos fonemas nasais não estão representados no inglês. Do ponto de vista prático, o Piper emerge como a opção mais adequada para produção de bancos fonéticos estáveis no curto prazo, enquanto StyleTTS 2 e Kokoro representam a aposta mais promissora à medida que o suporte ao PT-BR amadurecer.

6. Conclusões

Este trabalho contribuiu com três achados principais: uma avaliação comparativa de cinco paradigmas de síntese por IA aplicados à voz infantil em PT-BR com análise de adequação ao método fônico, dimensão ausente na literatura; a demonstração de que o dataset multilingual-phonemes-10k-alpha incorporado ao StyleTTS 2 produz melhoria consistente na cobertura fonética do PT-BR, resultado reproduzível com baixo custo; e o diagnóstico de que os desafios de síntese emocional e fonêmica são estruturais ao campo, o que reorienta o esforço de pesquisa de migração de plataforma para construção de corpora anotados.

Em termos práticos: ElevenLabs é adequado apenas para prototipagem neutra rápida; XTTS-v2 oferece clonagem local de qualidade quando há tempo de GPU disponível; Piper TTS é a melhor opção atual para bancos fonéticos estáveis, pela reprodutibilidade e inferência sem GPU; StyleTTS 2 tem maior potencial para PT-BR no médio prazo; Kokoro-82M é eficiente quando o tempo de treinamento é crítico; F5-TTS com RVC é indicado apenas quando entrada humana ao conversor é viável e a consistência tímbrica entre locutores é aceitável.

7. Limitações

A avaliação foi conduzida por consenso qualitativo de equipe multidisciplinar, sem protocolo MOS formal, avaliação cega ou sistematização de taxa de aprovação por ferramenta, percentual de erro fonêmico ou escala numérica de avaliação emocional. Em particular, a melhoria fonética atribuída à incorporação do dataset multilingual-phonemes-10k-alpha foi observada por inspeção qualitativa da equipe, sem ablação controlada ou medições objetivas de antes/depois; trata-se de uma observação promissora que ainda carece de validação estatística. O corpus de treinamento está abaixo do ideal para qualquer um dos modelos avaliados; o looping mitiga parcialmente essa restrição, mas pode favorecer overfitting em vez de ampliar genuinamente a cobertura fonética. Todas as avaliações foram realizadas em um único contexto de aplicação; generalizações exigiriam validação independente.

8. Considerações Éticas

A clonagem da voz de um menor foi conduzida com aprovação do Comitê de Ética em Pesquisa com Seres Humanos da Universidade Federal de Mato Grosso (UFMT) e consentimento livre e esclarecido do locutor e de seus responsáveis legais. Todo processamento foi realizado localmente, sem transmissão de dados biométricos a servidores externos, em conformidade com a LGPD (com exceção dos testes estritamente necessários e anônimos na API comercial avaliada, conduzidos sob termos de serviço e privacidade adequados).

9. Declaração de uso de IA Generativa

Durante a condução e redação deste trabalho, ferramentas de Inteligência Artificial Generativa baseadas em Grandes Modelos de Linguagem (LLMs) foram empregadas de forma delimitada e exclusiva como ferramentas auxiliares para a revisão ortográfica e gramatical e para a padronização e organização das referências bibliográficas. O planejamento experimental, a execução metodológica, a análise crítica dos resultados comparativos e toda a autoria intelectual subjacente às ideias apresentadas neste manuscrito são de responsabilidade exclusiva dos autores humanos. O texto não contém dados empíricos forjados ou resultados sintetizados artificialmente por sistemas de IA.

Referências

Araújo, G. E. et al. (2026). Certas Palavras: A 1980s–90s Brazilian Radio Corpus to Test TTS Models in Noisy Multi-Speaker Dialogues. In Proceedings of PROPOR 2026. ACL Anthology.

Barakat, H.; Turk, O.; Demiroglu, C. (2024). Deep learning-based expressive speech synthesis: a systematic review of approaches, challenges, and resources. EURASIP

Journal on Audio, Speech, and Music Processing. DOI: 10.1186/s13636-024-00329-7.

Bott, L. et al. (2024). Controlling Emotion in Text-to-Speech with Natural Language Prompts. In Proceedings of Interspeech 2024. ISCA.

Casanova, E.; Júnior, A. C.; Shulby, C.; Oliveira, F. S.; Teixeira, J. P.; Ponti, M. A.; Aluísio, S. (2022). TTS-Portuguese Corpus: a corpus for speech synthesis in Brazilian Portuguese. Language Resources and Evaluation, 56, 1043–1055. DOI: 10.1007/s10579-021-09570-4.

Casanova, E. et al. (2022b). YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for Everyone. In Proceedings of ICML 2022. PMLR.

Casanova, E. et al. (2024). XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model. arXiv:2406.04904.

Casanova, E.; Santos, V. G.; Svartman, F. R. F.; Leite, M. Q.; Candido Junior, A.; Marcacini, R. M.; Rezende, S. O.; Aluísio, S. M. (2026). Recursos para o processamento de fala. In Caseli, H. M.; Nunes, M. G. V. (orgs), Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português. 4ª ed., v.1. BPLN. DOI: 10.5281/zenodo.19259398.

Chen, X. et al. (2024). F5-TTS: A Fairytaler that Fakes Fluent and Faithful Speech with Flow Matching. arXiv:2410.06885.

Cho, J. et al. (2025). ECE-TTS: A Zero-Shot Emotion Text-to-Speech Model with Simplified and Precise Control. Applied Sciences, MDPI, 15(9), 5108.

ElevenLabs. (2024). Voice cloning and text-to-speech API. Disponível em: elevenlabs.io.

Gomes, E.; Pedroso, F. S.; Wagner, M. B. (2008). Hipersensibilidade auditiva no transtorno do espectro autístico. Pró-Fono, v. 20, p. 279–284.

Jurafsky, D.; Martin, J. H. (2023). Speech and Language Processing (3rd ed. draft). Stanford University.

Li, Y. A. et al. (2023). StyleTTS 2: Towards Human-Level Text-to-Speech through Style Diffusion and Adversarial Training with Large Speech Language Models. In Proceedings of NeurIPS 2023. arXiv:2306.07691.

Maia, R. et al. (2023). Prosódia e síntese da fala: uma revisão integrativa da literatura em português brasileiro. Revista da ABRALIN.

Oliveira, J.; Albuquerque, F. E. (2021). Leitura e escrita em crianças com autismo: o trabalho psicopedagógico a partir do método fônico. Facit Business and Technology Journal, 1(29).

Otto-Meyer, S. et al. (2018). Children with autism spectrum disorder have unstable neural responses to sound. Experimental Brain Research, 236, 733–743.

Rebouças, G. et al. (2021). Modelagem 3D do Robô Otto para o atendimento de crianças com Transtorno do Espectro Autista. In: Anais da XXI Escola Regional de Informática de Mato Grosso (ERI-MT), p. 35-41. Porto Alegre: SBC.

Scassellati, B.; Admoni, H.; Mataric, M. (2012). Robots for Use in Autism Research. Annual Review of Biomedical Engineering, v. 14, p. 275–294.

StyleTTS2 Community. (2024). multilingual-phonemes-10k-alpha. Hugging Face Datasets. Disponível em: huggingface.co/datasets/styletts2-community/multilingual-phonemes-10k-alpha.

[Software] Chen, X. et al. (2024). F5-TTS. GitHub: SWivid/F5-TTS. Disponível em: github.com/SWivid/F5-TTS.

[Software] Hexgrad. (2024). Kokoro-82M. Hugging Face. Apache 2.0. Disponível em: huggingface.co/hexgrad/Kokoro-82M.

[Software] OHF-Voice. (2025). Piper 1 — Fast and local neural text-to-speech engine. GitHub: OHF-Voice/piper1-gpl. GPL.

[Software] RVC Project. (2023). Retrieval-based Voice Conversion WebUI. GitHub: RVC-Project/Retrieval-based-Voice-Conversion-WebUI.

Evaluating LLM-based Triple Extraction for Knowledge Graph Fact-Checking in Portuguese

Roney Lira de Sales Santos[1], **Lucas dos Santos**[1], **João Pedro Holanda Souza**[2]

[1]Federal University of Bahia (UFBA)
Camaçari – BA – Brazil

[2]Universidade Norte do Paraná (UNOPAR)
Londrina – PR – Brazil

roneysantos@ufba.br, lukas.santtos53@gmail.com, joaocmdt505@gmail.com

Abstract. *Automatic fake news detection in Portuguese is still often treated as a text classification task, without explicitly representing the factual veracity of the claims. In this work, we evaluate LLM-based triple extraction in a knowledge graph (KG)-based fact-checking system. We replace the Open Information Extraction component of a previous approach with triples generated by Sabiá 4, while keeping the remaining pipeline unchanged. The graph is built only from triples extracted from true news articles and is used as factual support to verify new instances. The experiments use true and fake news articles across four evaluation settings. In the closed setting with the complete KG, LLM-based extraction achieves an F1 score of 0.9992, outperforming the OIE-based configuration. However, in more restrictive evaluation settings, expressive triples require entity normalization, relation standardization, and semantic consolidation to provide factual support beyond direct evidence.*

1. Introduction

The spread of misinformation in digital environments has motivated the development of automatic methods for fake news detection. Although supervised approaches achieve strong results, many of them rely on linguistic and statistical cues from the text, without explicitly modeling the factual veracity of the content [Vosoughi et al. 2018].

Content-based methods can address this limitation by representing claims as semantic relations in knowledge graphs (KGs) [Ciampaglia et al. 2015]. In this context, the work of [Santos et al. 2026] proposed a system for Portuguese in which news articles are decomposed into triples in the form of subject, relation[1], and object. Triples extracted from true news articles are used to build a KG, which serves as a structured source of factual support for evaluating new instances.

In this approach, triple extraction is a critical step. In [Santos et al. 2026], this process is performed with Open Information Extraction (OIE) [Banko et al. 2007], a technique with known limitations in Portuguese, including difficulties with implicit subjects, complex syntactic structures, and relations distributed across the text [Stanovsky and Dagan 2018]. Errors at this stage affect which facts are represented in the KG and, consequently, the veracity scores assigned by the system.

[1]We use *relation* to refer to the textual predicate extracted from the news article. This choice emphasizes that relations are open textual expressions generated by the extractor, rather than predicates from a fixed schema.

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 361–374
October 19-22, 2026 ©2026 Association for Computational Linguistics

Large language models (LLMs) can offer a promising alternative for this step, with the potential to extract relations with greater contextual coherence and better handle long-range dependencies [OpenAI et al. 2024]. However, it remains unclear how this substitution affects KG-based factual verification systems, especially for news in Portuguese.

In this work, we evaluate whether triple extraction with LLMs strengthens a KG-based fake news detection system for Portuguese. To this end, we take the approach of [Santos et al. 2026] as reference, replacing its OIE component with triples generated by an LLM while keeping the remaining pipeline components unchanged. This setting allows us to analyze, in a controlled way, how changes in factual representation affect the graph structure, the retrieval of factual support, and veracity inference.

The results show that LLM-based extraction outperforms the OIE-based approach in the closed setting with the complete KG, achieving an F1 score of 0.9992 compared with 0.9396. In more restrictive evaluations, such as *leave-one-news-out*, *leave-one-triple-out*, and train/test split, the system behaves conservatively when direct evidence is removed or when the evaluated facts fall outside the KG coverage. This result indicates that more expressive triples do not, by themselves, guarantee greater factual support beyond the evidence directly represented in the graph.

The main contributions of this work are: (i) the evaluation of an LLM as a triple extractor in a KG-based fact-checking pipeline for Portuguese; (ii) a controlled comparison with an OIE-based approach in the closed evaluation setting; and (iii) an analysis of the impact of triple extraction on graph structure, retrieval of factual support, and the veracity scores assigned to news articles.

The remainder of this paper is organized as follows. Section 2 reviews related work on fake news detection, fact-checking, and triple extraction. Section 3 describes the datasets, extraction procedure, KG construction, truth-value computation, and evaluation protocols. Section 4 presents and discusses the results, and Section 5 concludes the paper and outlines directions for future work.

2. Related Work

Automatic fake news detection in Portuguese has been studied mainly through linguistic cues, stylometric features, and supervised models. The Fake.Br Corpus [Monteiro et al. 2018, Silva et al. 2020] advanced this line of research by providing true and fake news articles aligned by topic, enabling controlled experiments for Portuguese. Broader reviews and surveys of the area report strategies based on linguistic attributes, machine learning, textual structure, and content-based methods [Santos 2022]. Although relevant, these approaches tend to classify the news article as a textual unit, without explicitly representing the events described in it as verifiable facts.

In automatic fact-checking, the work of [Ciampaglia et al. 2015] formulates factual verification as a connectivity problem in knowledge graphs. In this approach, a claim is represented as a triple in the form of subject, relation, and object, and its truth value is estimated from the semantic proximity between entities in a graph derived from DBpedia [Auer et al. 2007]. In a complementary direction, [Thorne et al. 2018] consolidated tasks in which claims are classified based on evidence retrieved from external sources. Recent surveys show that automatic fact-checking combines claim extraction, evidence retrieval, knowledge representation, and inference [Guo et al. 2022].

For Portuguese, the work of [Santos and Pardo 2020] adapted knowledge graph-based verification to resources derived from Portuguese DBpedia and also explored strategies based on search engine results. More recently, [Santos et al. 2026] proposed a graph-based system built from true news articles in Portuguese. Unlike encyclopedic graphs, this method represents news articles as factual events extracted into triples and computes the veracity of new instances from the support found in the graph. Since this step depends on Open Information Extraction (OIE), a technique designed to extract relations without a predefined schema [Banko et al. 2007], its quality depends strongly on syntactic analysis and on how well linguistic resources fit the target language. For this reason, extraction in Portuguese is identified as one of the main bottlenecks of this approach.

More recently, LLMs have been explored in information extraction and knowledge graph construction. Some studies review their use in generative extraction of entities, relations, and events [Xu et al. 2024], while others investigate strategies for triple extraction and graph construction from free text [Van Cauter et al. 2024, Mihindukulasooriya et al. 2025]. These works suggest that LLMs can reduce some limitations of traditional extractors, but they also point to challenges related to standardization, redundancy, hallucination, and consistency in the extracted triples.

This work differs from previous studies by evaluating the impact of LLM-based triple extraction in a knowledge graph-based factual verification approach for Portuguese, considering final performance, graph structure, retrieval of factual support, and the truth values assigned to news articles.

3. Methodology

This section describes the methodology used to evaluate the impact of LLM-based triple extraction on a knowledge graph-based fake news detection system. In this paper, we use KG, from *Knowledge Graph*, to refer to the graph built from triples extracted from true news articles. The approach follows the principle of factual support verification in graphs adopted in [Santos et al. 2026], but replaces OIE-based extraction with a large language model, while keeping the remaining pipeline components unchanged.

3.1. Approach Overview

The pipeline consists of five steps: (i) consolidation of the news datasets; (ii) extraction of triples structured as `(subject, relation, object)`; (iii) construction of the KG with triples from true news articles; (iv) computation of the truth value of triples extracted from new articles; and (v) final classification based on the aggregation of the obtained values.

Each news article is represented as a set of triples in the form of subject, relation, and object. The central hypothesis is that the KG works as a structured memory of previously observed factual events and relations. Thus, the larger and more representative the KG is, the greater its expected capacity to retrieve factual support for new articles. Low values indicate absence or insufficiency of retrievable support in the available KG, rather than absolute proof of falsity.

3.2. Datasets

The experiments use three public datasets for fake news detection in Portuguese: Fake.Br, FakeRecogna [Garcia et al. 2022], and FakeTrueBR [Chavarro et al. 2023].

These datasets were consolidated into a single collection of 22,684 news articles, balanced between true and fake instances.

Fake.Br and FakeRecogna were used in already preprocessed versions, with texts normalized through steps such as tokenization, lemmatization, and/or stemming. We then standardized the fields and mapped the labels of the three datasets to a common binary representation, in which fake news articles receive label 0 and true news articles receive label 1. Table 1 summarizes the data.

Dataset	True	Fake	Total
Fake.Br	3,600	3,600	7,200
FakeRecogna	5,951	5,951	11,902
FakeTrueBR	1,791	1,791	3,582
Total	11,342	11,342	22,684

Table 1. Composition of the datasets used in the experiments.

3.3. LLM-based Triple Extraction

Triple extraction was performed with Sabiá 4 [Laitz et al. 2026], a large language model developed by Maritaca AI[2] with a focus on Brazilian Portuguese. This choice is motivated by its alignment with the evaluated corpora, all of which are in Brazilian Portuguese. According to [Laitz et al. 2026], Sabiá 4 underwent continual pretraining on Portuguese and Brazilian legal corpora, supports a context window of 128 thousand tokens, and was tuned with instruction and preference data.

Each news article was submitted to the model with a fixed instruction to extract triples in the form of subject, relation, and object. The prompt was designed to reduce hallucinations, avoid unsupported inferences, and control the granularity of the relations. The model was instructed to extract only explicit information, produce a single semantic relation per triple, decompose objects containing embedded clauses or relations, preserve modality and speech attribution when relevant, and return only valid JSON in the format `[["subject", "relation", "object"], ...]`.

The calls were made through the Maritaca AI API[3], using a configuration compatible with the OpenAI Chat Completions API[4]. We used temperature 0 and a maximum limit of 1,000 tokens per response. Processing was executed in parallel, with request-per-minute control and retries in case of failure. The responses were cleaned and validated to remove Markdown markers and identify the list of triples. We consider a triple valid when it is extracted in a structured format with exactly three non-empty textual elements, corresponding to subject, relation, and object.

3.4. Knowledge Graph Construction

The KG was built from the JSONL file generated during extraction. Each record contains the news article identifier, its label, and a list of triples in the form of subject, relation,

[2]Available at: `https://www.maritaca.ai/`. Accessed: 12 May 2026.

[3]Available at: `https://docs.maritaca.ai/pt/visao-geral`. Accessed: 12 May 2026.

[4]Available at: `https://developers.openai.com/api/reference/chat-completions/overview`. Accessed: 12 May 2026.

and object. Before insertion, each triple was validated to ensure three non-empty textual elements.

Only triples extracted from true news articles were used to build the KG. Thus, the graph works as a structured memory of events and relations observed in news articles considered reliable, while fake news articles are used only for evaluation.

Formally, $KG = (V, E)$, where V represents subjects and objects, and E represents relations between them. Each triple (s, r, o) adds the nodes s and o to the graph, when not already present, and an edge labeled with relation r.

The implementation was carried out with NetworkX [Hagberg et al. 2008]. In the main experiments, we use an undirected multigraph, `nx.MultiGraph()`, allowing multiple relations between the same pair of entities. The choice of an undirected version follows the semantic proximity formulation based on paths in knowledge graphs [Ciampaglia et al. 2015].

3.5. Truth Value Computation

To verify a news article, the method computes a truth value for each triple based on the support found in the KG. Let $t = (s, r, o)$ be a triple. The value $\tau(t) \in [0, 1]$ is estimated from the connectivity between s and o, following the semantic proximity intuition of [Ciampaglia et al. 2015]. The relation r is preserved as the edge label, but the computation considers only the connectivity between subject and object.

If s and o correspond to the same node, or if there is a direct edge between them, we assign $\tau(t) = 1$. If either node is absent from the graph, or if there is no path between them, we assign $\tau(t) = 0$. In all other cases, the shortest weighted path between s and o is computed with Dijkstra's algorithm[5] [Dijkstra 1959]. The path cost is the sum of $\log(k(v))$ over the intermediate nodes, where $k(v)$ is the degree of node v considering relation multiplicity. The triple value is defined by Equation 1 [Ciampaglia et al. 2015]:

$$\tau(t) = \frac{1}{1 + \sum_{v \in P_{s,o}} \log(k(v))} \tag{1}$$

where $P_{s,o}$ contains the intermediate nodes in the path between s and o. Thus, direct connections or paths mediated by specific nodes receive higher values, while long or less informative paths receive lower support.

The final value of a news article N is computed as the mean truth value of its triples, as shown in Equation 2 [Santos 2022, Santos et al. 2026]:

$$Score(N) = \frac{1}{|T_N|} \sum_{t \in T_N} \tau(t) \tag{2}$$

where T_N is the set of triples extracted from the news article. News articles with no valid triples receive $Score(N) = 0$, since they provide no retrievable factual support in the KG.

Classification follows three decision ranges: $Score(N) \geq 0.60$ indicates a true news article; $Score(N) < 0.40$ indicates a fake news article; and values in $[0.40, 0.60)$ indicate an uncertain case, as in [Santos et al. 2026]. For binary comparisons, the uncertain class is conservatively mapped to fake.

[5]Dijkstra's algorithm was chosen because it computes shortest paths in weighted graphs with non-negative weights, which is the case for the cost function based on logarithmic node degrees.

3.6. Evaluation Protocols

The evaluation was conducted in four settings:

- **Complete KG**: the KG is built with all triples extracted from true news articles, and verification is applied to the full dataset. This is a closed evaluation setting because true news articles evaluated by the system may also have contributed triples to the KG. Therefore, this setting measures the internal compatibility between articles and graph, indicating the maximum factual retrieval capacity when facts are represented in the KG.
- ***Leave-one-news-out***: when evaluating a true news article, all its triples are temporarily removed from the KG. This setting reduces the effect of direct memorization and tests whether other true news articles provide factual support for the evaluated events.
- ***Leave-one-triple-out***: the evaluated triple is removed from the KG before computing its truth value, when it appears as a direct connection. This protocol follows the intuition of removing direct evidence in knowledge graphs [Ciampaglia et al. 2015] and checks whether there is indirect support for the claim.
- **Train/test split**: the KG is built only with triples from true news articles in the training set, and evaluation is performed on the test set. We used an 80/20 split, empirically chosen to preserve most true factual evidence for KG construction while still reserving a representative portion of the data for evaluation. This is the setting closest to a real application, since it evaluates news articles not used to build the graph.

In addition to accuracy, precision, recall, F1, and AUC, we report *coverage* [Geifman and El-Yaniv 2017], defined as the proportion of news articles for which the system issues a true or fake decision. This measure is relevant because the method includes an uncertain class, used when retrievable factual support in the KG is insufficient. Thus, low support in the KG should be interpreted as absence of retrievable evidence in the available base, and not necessarily as absolute falsity.

4. Results and Discussion

This section presents the results of LLM-based triple extraction in the KG-based pipeline. The analysis covers the characterization of triples and the graph, the evaluation protocols, the comparison with OIE, and the main effects observed in factual retrieval.

4.1. Characterization of Triples and the Graph

Extraction with Sabiá 4 generated 296,362 valid triples from 22,684 news articles, with an average of 13.06 triples per article and a median of 11. In total, 1,170 articles had no valid triples, corresponding to 5.16% of the corpus. Table 2 summarizes the distribution of triples by class.

True news articles generated more triples on average, suggesting higher factual density in the texts used to build the KG. However, the presence of true articles with no valid triples indicates that extraction is still sensitive to textual structure and to how events are expressed.

Class	Articles	Triples	Mean	No triples
Fake	11,342	116,601	10.28	179
True	11,342	179,761	15.85	991
Total	22,684	296,362	13.06	1,170

Table 2. Statistics of the triples extracted by Sabiá 4.

The KG built from true news articles contains 253,464 nodes and 217,485 edges when multiplicity is considered, of which 209,188 are unique edges. The structure has 52,480 connected components, with the largest component containing 112,588 nodes, equivalent to 44.42% of the total. Table 3 presents the main structural statistics.

Metric	Value
Nodes	253,464
Edges with multiplicity	217,485
Unique edges	209,188
Connected components	52,480
Nodes in largest component	112,588
Largest component proportion	44.42%
Average degree	1.72
Nodes with degree 1	218,622
Distinct relations	76,681

Table 3. Structural statistics of the KG built from triples extracted from true news articles.

These results indicate that LLM-based extraction produces a broad and semantically varied graph, but also a sparse one. The large number of nodes with degree 1, corresponding to 86.25% of the nodes, shows that many entities appear in low-recurrence relations. Figure 1 reinforces this pattern, with a distribution concentrated in low degrees and a long tail formed by a small number of highly connected nodes. For factual verification, this means that the KG stores many events, but part of them remains in poorly connected regions.

4.2. Truth Value Evaluations

The four evaluations were designed to verify whether the KG retrieves facts present in the graph itself, whether the system depends on the direct presence of the evaluated article, whether indirect support exists when the corresponding edge is removed, and whether the KG evaluates unseen news articles from a previously built factual base. Table 4 presents the results of the four protocols. For binary comparison, instances classified as uncertain were mapped to the fake class, following a conservative policy.

In the complete KG setting, the system reaches an F1 score of 0.9992 and an AUC of 0.9993, indicating strong factual retrieval when the events from true news articles are represented in the graph. This result, however, should be interpreted with caution: since the KG is built with all available true news articles, part of the evaluated true articles also contributed to the construction of the graph itself. Therefore, this protocol mainly measures the internal consistency of the KG and its ability to retrieve facts already present in the base, rather than the system's performance on completely new articles.

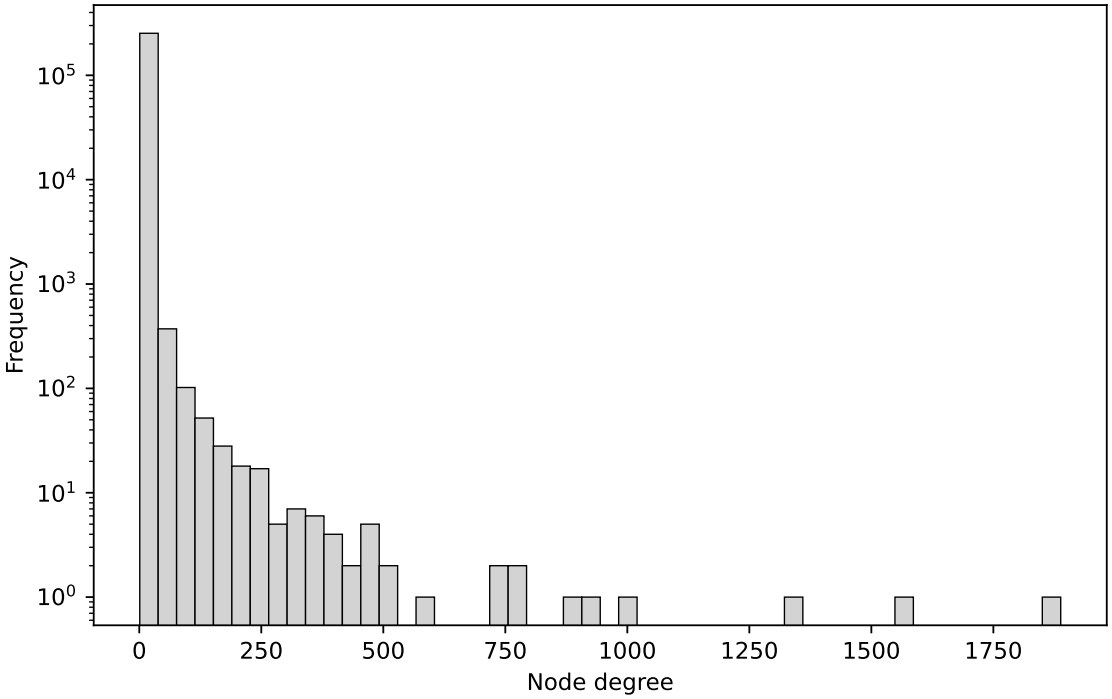


Figure 1. Degree distribution of the nodes in the KG built from triples extracted by Sabiá 4.

Protocol	Acc.	Prec.	Rec.	F1	AUC	Coverage
Complete KG	0.9992	0.9993	0.9991	0.9992	0.9993	0.9989
Leave-one-news-out	0.5201	0.9831	0.0410	0.0787	0.5475	0.9927
Leave-one-triple-out	0.5202	0.9832	0.0412	0.0790	0.5832	0.9919
Train/test split	0.5147	1.0000	0.0291	0.0566	0.5375	0.9938

Table 4. Results of the evaluation protocols with triples extracted by Sabiá 4.

In the restrictive[6] settings, the behavior changes substantially. The F1 score drops to 0.0787 in *leave-one-news-out*, 0.0790 in *leave-one-triple-out*, and 0.0566 in the train/test split. In all cases, precision remains high, but recall is very low. This indicates that the system avoids classifying an article as true without strong retrievable evidence, but fails to recognize many true articles when their events are not sufficiently connected to the KG.

Figure 2 confirms this behavior. In the complete KG setting, the means clearly separate fake and true news articles. In the restrictive protocols, the values of true articles shift to low ranges, approaching the values of fake articles. Thus, the performance drop is associated with KG coverage and connectivity, and not necessarily with the absence of factual differences between classes.

[6]We call these protocols restrictive because, unlike the complete KG setting, they reduce the evidence available for truth value computation by removing the article itself, the evaluated triple, or by splitting the data into training and test sets.

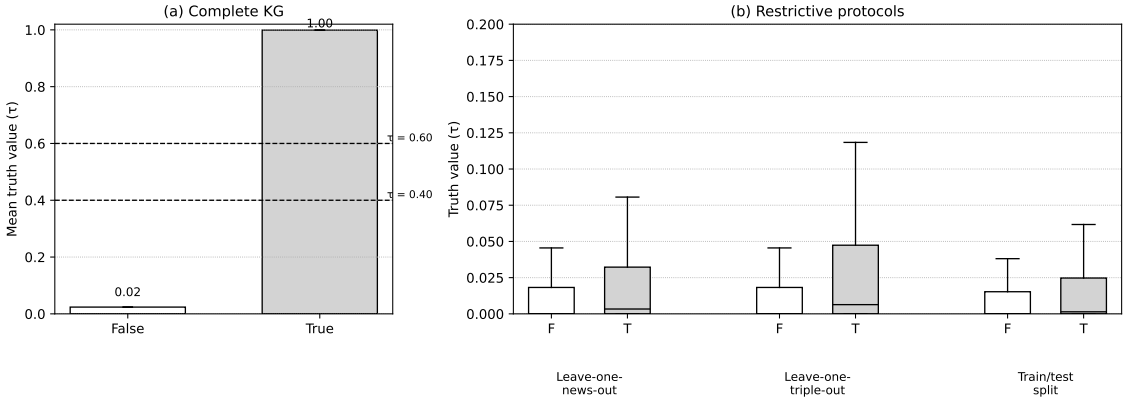


Figure 2. Truth values by class and protocol. Panel (a) shows the mean value in the complete KG setting, while panel (b) presents the distribution of values in the restrictive protocols.

4.3. Comparison with OIE

Table 5 compares extraction with Sabiá 4 to the OIE-based system reported in [Santos et al. 2026]. The comparison is made only in the complete KG setting, since this is the closest to the configuration presented in the reference work; the other protocols have no direct equivalent.

Extraction	Acc.	Prec.	Rec.	F1
OIE [Santos et al. 2026]	0.9359	0.8887	0.9966	0.9396
LLM (Sabiá 4)	0.9992	0.9993	0.9991	0.9992

Table 5. Comparison between OIE and LLM (Sabiá 4) in the closed evaluation setting.

Because the OIE results are available only for the complete KG setting, this comparison supports conclusions only about factual retrieval in the closed setting. It does not establish that Sabiá 4 is more robust than OIE when direct evidence is removed or when previously unseen articles are evaluated. Reproducing the restrictive protocols with OIE-based extraction is therefore necessary for a broader comparison.

4.4. Discussion and Qualitative Analysis

The results should be interpreted in light of the role of the KG in the method. The graph works as a structured memory of events and relations extracted from true news articles. Thus, the score expresses the degree of factual support available in the KG, rather than absolute proof of truth or falsity. When factual evidence is removed or was never inserted into the graph, many true articles receive low truth values due to lack of retrievable support.

In a practical deployment, this factual memory would require continuous maintenance. Newly verified news articles could be periodically ingested with source provenance and temporal metadata attached to each triple. Such metadata would help distinguish current evidence from outdated or conflicting information and support the revision or removal of facts that are no longer valid.

The qualitative analysis of triples helps explain this behavior. Sabiá 4 preserves information relevant to journalistic texts, such as speech attribution, modality, and contextual relations, distinguishing statements such as *"o ministro afirmou que a medida será revista"* (the minister stated that the measure will be reviewed), *"o ministro negou a medida"* (the minister denied the measure), and *"a medida foi revista"* (the measure was reviewed). At the same time, we observed long subjects, objects, and relations, as well as textual variation for semantically close entities. These phenomena enrich the representation, but also fragment the graph.

For example, forms such as *"presidente Lula"* (President Lula), *"Lula"*, and *"Luiz Inácio Lula da Silva"* may refer to the same entity, while relations such as *"informou"* (reported), *"afirmou"* (stated), *"declarou"* (declared), and *"disse"* (said) may play similar discourse roles. Since they appear as distinct nodes and relations in the KG, they reduce recurrence and make it harder to form paths between related facts.

Thus, the comparison with OIE indicates that LLMs improve factual retrieval in the closed setting. However, the restrictive protocols reveal a broader limitation of KG-based approaches: their dependence on graph coverage and connectivity. This limitation may affect triples extracted by both LLMs and OIE. In the case of LLMs, the greater textual richness of the triples may increase fragmentation when there is no entity and relation normalization.

5. Conclusion

This paper evaluated the use of LLMs for triple extraction in a knowledge graph-based fake news detection system for Portuguese. Starting from a previously proposed approach, we replaced the OIE-based extraction step with triples generated by Sabiá 4, while keeping the remaining pipeline components unchanged. We then analyzed how the factual representation extracted from news articles affects KG construction, factual support retrieval, and veracity inference.

The results indicate that the approach is promising for factual verification in Portuguese when the relevant events are represented in the KG. In the complete KG setting, LLM-based extraction provides strong factual retrieval and outperforms the OIE-based configuration in the closed evaluation setting. In the restrictive protocols, however, the greater expressiveness of the triples does not guarantee greater factual support when direct evidence is removed or when the evaluated facts fall outside the KG coverage. This behavior may also occur with OIE and should be investigated in future studies that replicate the same protocols with different extractors.

These findings reinforce that extraction is a decisive step in graph-based fact-checking systems, as it determines which facts are stored, how they are connected, and which evidence can be retrieved. Future work should compare LLMs from different families and sizes, including standard instruction-following and reasoning-oriented configurations, under the same extraction and evaluation protocols. We also plan to investigate entity and relation normalization, coreference resolution, entity disambiguation, and the use of ontologies or lightweight semantic schemas to reduce KG fragmentation. In addition, the graph should be continuously expanded with newly verified news articles, while preserving source provenance and temporal metadata to distinguish current evidence from outdated or conflicting information. Finally, we intend to explicitly incorporate the rela-

tion into the truth value computation, allowing the method to distinguish triples with the same pair of entities but different relations, such as *"o ministro anunciou a medida"* (the minister announced the measure) and *"o ministro negou a medida"* (the minister denied the measure).

Limitations

This work has some limitations. First, the truth value computed by the method should be interpreted as the degree of factual support retrievable from the available KG, rather than as absolute proof of truth or falsity. Thus, true news articles whose facts are not yet represented in the graph may receive low support. Second, although extracted relations are stored as edge labels, the current computation mainly considers connectivity between subject and object. This may assign similar support to triples with the same pair of entities but different relations. Third, LLM-based extraction may generate textual variation, redundancy, or long relations, contributing to KG fragmentation. Fourth, the experiments use a single LLM, Sabiá 4. Therefore, the results do not isolate the effects of model family, size, training data, or reasoning-oriented inference and should not be generalized to LLM-based extractors as a whole. Finally, the restrictive protocols were applied to the LLM-based configuration, but still need to be reproduced with OIE-based configurations for a broader comparison.

Ethics Statement

This work uses only public datasets already made available for research on fake news detection in Portuguese. We did not collect personal data directly from users, nor did we conduct interventions with human participants. The news articles are handled in aggregate form and used exclusively for experimental evaluation. We acknowledge, however, that automatic misinformation detection systems may produce errors, especially when factual coverage in the KG is limited. Therefore, the method's outputs should be interpreted as support for human analysis, not as definitive decisions about the veracity of a news article.

Generative AI Use

Generative AI tools were used to support textual revision of the manuscript, including grammatical and spelling corrections, and limited assistance with LaTeX formatting, such as emphasis, tables, and section structure. Generative AI was not used to produce experimental results, modify data, generate metrics, or replace the authors' scientific analysis. All methodological decisions, interpretations of results, and final versions of the text were reviewed and validated by the authors.

Acknowledgments

This research was fully supported by the Federal University of Bahia (UFBA) through the JovemPesq Program, under Call for Proposals PRPPG No. 023/2025. The authors also gratefully acknowledge the financial support provided by the Brazilian National Council for Scientific and Technological Development (CNPq).

References

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). DBpedia: A nucleus for a web of open data. In *The Semantic Web*, volume 4825 of *Lecture Notes in Computer Science*, pages 722–735, Berlin, Heidelberg. Springer.

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., and Etzioni, O. (2007). Open information extraction from the web. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 2670–2676. Morgan Kaufmann Publishers Inc.

Chavarro, J. P., Carvalho, D. A., and Rezende, S. O. (2023). FakeTrueBR: Um corpus brasileiro de notícias falsas. In *Anais da XVIII Escola Regional de Banco de Dados*, pages 53–62, Porto Alegre, RS, Brasil. SBC.

Ciampaglia, G. L., Shiralkar, P., Rocha, L. M., Bollen, J., Menczer, F., and Flammini, A. (2015). Computational fact checking from knowledge networks. *PLOS ONE*, 10(6):e0128193.

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271.

Garcia, G. L., Afonso, L. C. S., and Papa, J. P. (2022). FakeRecogna: A new brazilian corpus for fake news detection. In Pinheiro, V., Gamallo, P., Amaro, R., Scarton, C., Batista, F., Silva, D., Magro, C., and Pinto, H., editors, *Computational Processing of the Portuguese Language*, volume 13208 of *Lecture Notes in Computer Science*, pages 57–67, Cham. Springer.

Geifman, Y. and El-Yaniv, R. (2017). Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30.

Guo, Z., Schlichtkrull, M., and Vlachos, A. (2022). A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference*, pages 11–15.

Laitz, T., Almeida, T. S., Abonizio, H., Malaquias Junior, R., Bonás, G. K., Piau, M., Larcher, C., Pires, R., and Nogueira, R. (2026). Sabiá-4 technical report.

Mihindukulasooriya, N., Tiwari, A., Galkin, M., Costabello, L., and Simperl, E. (2025). Automatic prompt optimization for knowledge graph construction with large language models. In *Proceedings of the Workshops of the VLDB 2025 Conference*.

Monteiro, R. A., Santos, R. L. S., Pardo, T. A. S., Almeida, T. A., Ruiz, E. E. S., and Vale, O. A. (2018). Contributions to the study of fake news in portuguese: New corpus and automatic detection results. In *Computational Processing of the Portuguese Language*, volume 11122 of *Lecture Notes in Computer Science*, pages 324–334, Cham. Springer.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R.,

Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. (2024). Gpt-4 technical report.

Santos, L. d., Santos, M. R. E., Souza, Y. S., Sousa, J. P. H., and Santos, R. L. d. S. (2026). Exploring knowledge graphs for automatic fake news detection in Portuguese. In *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 446–455, Salvador, Brazil. Association for Computational Linguistics.

Santos, R. L. d. S. (2022). *Detecção automática de notícias falsas em português*. Tese (doutorado em ciências – ciências de computação e matemática computacional), Universidade de São Paulo, São Carlos.

Santos, R. L. d. S. and Pardo, T. A. S. (2020). Fact-checking for portuguese: Knowledge graph and google search-based methods. In Quaresma, P., Vieira, R., Aluísio, S.,

Moniz, H., Batista, F., and Gonçalves, T., editors, *Computational Processing of the Portuguese Language*, volume 12037 of *Lecture Notes in Computer Science*, pages 195–205, Cham. Springer.

Silva, R. M., Santos, R. L., Almeida, T. A., and Pardo, T. A. (2020). Towards automatically filtering fake news in portuguese. *Expert Systems with Applications*, 146:113199.

Stanovsky, G. and Dagan, I. (2018). Supervised open information extraction. In *Proceedings of NAACL-HLT*, pages 885–895.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. (2018). FEVER: A large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Van Cauter, Z., Demeester, T., Vercruyssen, V., and Ongenae, F. (2024). Ontology-guided knowledge graph construction from maintenance short texts. In *Proceedings of the 1st Workshop on Knowledge Augmented Large Language Models*, pages 79–88. Association for Computational Linguistics.

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380):1146–1151.

Xu, D., Chen, W., Peng, W., Zhang, C., Xu, T., Zhao, X., Wu, X., Zheng, Y., and Wang, E. (2024). Large language models for generative information extraction: A survey. *Frontiers of Computer Science*, 18(6):186357.

Prompt Engineering for Small Language Models: Evaluating ICL for Portuguese Sentiment Analysis

André da F. Schuck[1], Gabriel L. Garcia[1], João Renato R. Manesco[1], Pedro Henrique Paiola[1], Leandro A. Passos [1], João Paulo Papa[1]

[1]São Paulo State University (UNESP) / Av. Eng. Luís Edmundo Carrijo Coube, 14-01 – Bauru – SP – Brazil

{andre.schuck, gabriel.lino, joao.r.manesco, pedro.paiola, leandro.passos, joao.papa}@unesp.br

Abstract. *The In-Context Learning (ICL) paradigm enables adapting LLMs without parameter tuning. This work evaluates the impact of prompt engineering on 9B models for Brazilian Portuguese sentiment classification, comparing six prompting formats (few- and zero-shot) across three linguistically diverse models (Gemma2-9B-it, Boto-9B-IT, Qwen3.5-9B) and four public datasets, anchored by a fine-tuned encoder (MSA-DistilBERT, ∼0.1B) and a 685B MoE model (DeepSeek-V3.2) under a unified protocol. Results show that all 9B models outperform the encoder and recover 71–101% of the weak-to-strong gap, with Qwen3.5-9B achieving the highest mean coverage (94.5%). Statistical tests indicate that well-defined instructions are the main driver of ICL performance, capturing most of the achievable accuracy even in zero-shot settings, while elaborate formats (roleplay, JSON schema, meta-prompting) add no consistent gain over a minimal instruction-plus-demonstrations prompt, offering practical guidance for deploying SLMs in resource-constrained PT-BR NLP scenarios.*

1. Introduction

The In-Context Learning (ICL) paradigm [Brown et al. 2020] enables Large Language Models (LLMs) to adapt to new tasks via natural language alone: rather than updating parameters, the model is conditioned on a small set of instructions and/or demonstrations supplied in its input context [Dong et al. 2024]. Prompt engineering, defined as the systematic design and optimization of such inputs [Schulhoff et al. 2024, Reynolds and McDonell 2021, White et al. 2023], has consolidated a flexible paradigm for task execution across numerous applied scenarios [Liu et al. 2021b, Zhao et al. 2024], including sentiment analysis (SA) [Xu et al. 2024]. ICL remains, however, notoriously brittle, since its performance can shift substantially with prompt format [Sclar et al. 2024] and demonstration ordering [Lu et al. 2022], motivating rigorous empirical study, particularly for underrepresented languages.

SA concerns the automatic identification of subjective orientation in text, typically as positive, negative, or neutral [Liu 2012, Pang and Lee 2008], and underpins applications ranging from market intelligence to public-health surveillance [Birjali et al. 2021]. For Brazilian Portuguese (PT-BR), SA remains challenging: opinion texts exhibit informality, slang, and culturally-specific idioms, while labeled corpora are scarce when compared to English [Pereira 2021].

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 375–389
October 19-22, 2026 ©2026 Association for Computational Linguistics

Despite the potential of ICL, systematic studies applying it to PT-BR remain scarce [Peres 2023, Assis et al. 2024], especially in SA [Marreira et al. 2025, Piorino et al. 2025, Lima et al. 2026].

This work addresses that gap by focusing on open-source Small Language Models (SLMs), defined here as Transformer models with at most 10 billion parameters [Oliveira et al. 2025], which can be deployed on modest local infrastructure while remaining competitive with massive proprietary LLMs [Lu et al. 2025, Nguyen et al. 2025, Wang et al. 2025]. Their availability raises a concrete deployment dilemma for PT-BR practitioners: under modest local infrastructure, should one fine-tune a small encoder, prompt an open 9B model, or pay for a frontier-scale API? Answering this requires more than leaderboard numbers, as it demands knowing how much performance depends on prompt design, how fragile each option is to formatting choices, and whether language-targeted adaptation helps or hurts.

Recent evidence suggests that SLMs are competitive for PT-BR SA [Lima et al. 2026], but also highly sensitive to prompt format and complexity [Schuck et al. 2025]; unlike those evaluations, which compare heterogeneous model families, we hold scale constant and isolate linguistic profile, including a controlled base→fine-tuned pair. We thus evaluate three open-source 9B SLMs: (i) Gemma2-9B-it [Gemma Team et al. 2024], representing English-dominant pretraining; (ii) Boto-9B-IT [Brígida 2024], a PT-BR-specialized LoRA fine-tune of Gemma2-9B-it; and (iii) Qwen3.5-9B [Qwen Team 2026], a natively multilingual model, anchored between a fine-tuned encoder floor (∼0.1B) and a 685B ceiling under a single unified protocol. The study is organized around three research questions:

- **RQ1:** Where do 9B models under ICL sit between the ∼0.1B encoder floor and the 685B ceiling, and does prompt sensitivity persist across scales?
- **RQ2:** Do instructions and demonstrations contribute distinctly to ICLperformance for PT-BR SA, and does combining them outperform either in isolation?
- **RQ3:** Do more elaborate combined formats yield statistically significant gains over the minimal one, or do they exhibit practical equivalence?

Our contributions are threefold: (i) a controlled PT-BR SA design that fixes scale at 9B and varies linguistic profile, including a base→fine-tuned pair (Gemma2/Boto) that isolates PT-BR adaptation at constant architecture and initialization, with the 685B baseline run under the same six prompts, demonstrations, and orderings so that prompt sensitivity, not only accuracy, becomes comparable across scales; (ii) evidence that 9B SLMs recover 71–101% of the weak-to-strong gap, with a natively multilingual model matching the 685B model's robustness; and (iii) evidence that well-defined instructions are the main driver of ICL performance, capturing most of the achievable accuracy even zero-shot, while demonstrations alone may degrade it [Min et al. 2022].

The remainder of this paper is organized as follows: Section 2 covers theoretical foundations; Section 3 details the experimental design; Section 4 reports results; Section 5 discusses findings; and Section 6 presents conclusions.

2. Background

Formally, ICL conditions the model's forward pass on a prompt p that concatenates input–output pairs (x_i, y_i), so prediction for a new x corresponds to the most probable continuation of $p \oplus x$ under the pre-training distribution f_θ [Brown et al. 2020, Wei et al. 2023, Dong et al. 2024]. As a research field, prompt engineering spans zero- and few-shot prompting [Brown et al. 2020], automated prompt generation [Shin et al. 2020, Zhou et al. 2023], demonstration selection [Liu et al. 2021a] and ordering [Lu et al. 2022], and reasoning-enhancing techniques [Wang et al. 2023].

3. Experimental Design

3.1. Tasks and Datasets

The proposed approaches were evaluated on two tasks, i.e., a binary sentiment classification (BSC: Positive/Negative) and a multiclass sentiment classification (MSC: Positive/Negative/Neutral) tasks, using four PT-BR datasets, detailed in Table 1. Datasets exceeding 1,500 test instances were subsampled via stratified random sampling with a fixed seed. Few-shot demonstrations were drawn exclusively from the training split; evaluation used the test split. Original class labels were mapped to numerical codes: -1 (Negative), 0 (Neutral), 1 (Positive).

Table 1. Datasets employed in the experimental evaluation. 🔴: `negative`; 🟡: `neutral`; and 🟢: `positive`. Distribution refers to the training split (source of demonstrations). Train and test splits share the same distribution.

Task	Dataset	Test Size	Class Distribution
BSC	IMDB_PT (IMDB)	1,500	🔴 50% ; 🟢 50%
	OPCovid-BR (OPC)	123	🔴 50% ; 🟢 50%
MSC	ReLI	1,500	🔴 4.82% ; 🟡 72.72% ; 🟢 22.46%
	TweetSentBr (TSBr)	1,500	🔴 33.3% ; 🟡 33.3% ; 🟢 33.3%

3.2. Models and Baselines

Table 2 summarizes the five evaluated systems, including three 9B-parameter SLMs to control for model size: Boto, a PT-BR-specialized variant of Gemma2 fine-tuned on the instruction-following dataset *cetacean-ptbr*[1], and Qwen3.5, which provides a multilingual contrast with support for 201 languages through native architectural design. The weak baseline (MSA-DistilBERT) represents the fine-tuning paradigm [Sanh et al. 2020, Devlin et al. 2018] and was used strictly off-the-shelf, with no dataset-specific adaptation: a multilingual encoder already fine-tuned for SA via supervised learning and evaluated without prompting. Its five-class output was consolidated by summing polarity extremes, i.e., $P(\text{very neg.}) + P(\text{neg.})$ and $P(\text{pos.}) + P(\text{very pos.})$, retaining Neutral for MSC and discarding it for BSC. Two asymmetries therefore constrain its interpretation: the model is tuned for the task but not for the target domains, and the BSC consolidation forces a polarity decision on instances whose modal class was Neutral. Both act in the same direction, so this

[1] `https://huggingface.co/datasets/lucianosb/cetacean-ptbr`

Table 2. Models and Baselines.

Role	Model Short	Parameters	Linguistic Profile
SLM	Gemma2[a]	9B	English-dominant
	Boto[b]	9B	PT-BR fine-tuned
	Qwen3.5[c]	9B	Multilingual (201 langs.)
Weak BL	MSA-DistilBERT[d]	0.1B	Multilingual (23 langs.)
Strong BL	DeepSeek[e]	685B	Multilingual

[a] `google/gemma-2-9b-it` [Gemma Team et al. 2024]
[b] `lucianosb/boto-9B-it` [Brígida 2024]
[c] `Qwen/Qwen3.5-9B` [Qwen Team 2026]
[d] `tabularisai/multilingual-sentiment-analysis` [TabularisAI et al. 2025]
[e] `deepseek-ai/DeepSeek-V3.2 (Non-thinking Mode)` [DeepSeek-AI 2025]

baseline should be read as a conservative floor and the gap-coverage percentages of Section 4.1 as correspondingly optimistic. The strong baseline (DeepSeek), in contrast, was evaluated under the same ICL protocol as the SLMs, ensuring comparability across tasks, datasets, prompt formats, demonstrations, and orderings.

3.3. Prompt Engineering

Table 3 presents the six evaluated prompt formats. *Demos-only* and the zero-shot *Inst-only* act as ablation anchors isolating the contributions of demonstrations and instructions (RQ2), and are competitive baselines in prior work [Ajith et al. 2023]; *Inst+Demos* is the simplest combined format, providing explicit classification guidance, against which three more elaborate variants are compared (RQ3): *JSON-Schema*, which enforces a predefined output schema for consistent parsing [Simmering and Huoviala 2023]; *Roleplay*, which assigns a domain-relevant persona [White et al. 2023, Zheng et al. 2024]; and *Meta-prompt*, an LLM-generated instruction (Claude 3.5 Sonnet, two refinement iterations) [Schulhoff et al. 2024].

All few-shot settings use 6 demonstrations (3 per class for BSC, 2 per class for MSC) in a "Q:{text}, A:{label}" format with English structural tokens and Portuguese content [Fu et al. 2022], selected via k-means clustering over *jina-embeddings-v3* representations to maximize semantic diversity [Sturua et al. 2024, Gao et al. 2023, Liu et al. 2021a], with three independent random orderings per configuration; full Jinja2 templates are available in the supplementary repository[2].

Table 3. Prompt format variants. All formats except *Inst-only* use 6 balanced demonstrations (3 per class for BSC, 2 per class for MSC) in the Q:A format. Full Jinja2 prompt templates are provided in the supplementary repository.

Variant	Components
Demos-only	Demonstrations only
Inst+Demos	Task instruction + demonstrations
JSON-Schema	Task instruction + JSON schema + demonstrations
Roleplay	Role specification + task instruction + demonstrations
Meta-prompt	Meta-prompt optimized instruction + demonstrations
Inst-only	Task instruction only (zero-shot)

[2] `https://github.com/AndreSchuck/slm_prompt_analysis`

3.4. Evaluation Methodology

Output parsing. Model outputs are parsed via format-specific regular expressions: for *JSON-Schema*, a single pattern targets the *sentimento* field via `re.search()`, tolerating common wrappers (e.g., Markdown fences, integer/float labels); for the remaining formats, a three-layer cascaded `re.match()` strategy in descending priority captures (i) a label at the start of the output, (ii) *"A resposta ... é {label}"*, and (iii) *"A classificação ... é {label}"*. Outputs unmatched by any layer are assigned code 2, denoting Format Non-Compliance (FNC). Full regex patterns and validation cases are available in the supplementary repository[2].

Metrics. Balanced Accuracy (BAcc), the mean of per-class Recall [Grandini et al. 2020], is the primary metric, essential given the 72.7% Neutral skew in ReLI. FNC outputs are retained rather than discarded and count as misclassifications, penalizing unparseable outputs identically to incorrect ones; their rate is reported separately as a diagnostic of output-structure adherence. Sensitivity to format is measured by the *Amplitude* (Δ), the BAcc range across the six prompt formats, and to ordering by the *Coefficient of Variation* (CV) [Brown 1998], σ/μ of BAcc across the three orderings, averaged over the five few-shot formats.

Prediction consolidation. Since the three iterations per configuration share the same inputs and differ only in the demonstration ordering, their predictions are not independent, and the consolidation is performed via majority voting [Dietterich 2000], yielding one prediction per instance that reflects the model's modal behavior. Tripartite ties, where each iteration predicts a different class (possible only in MSC), are assigned code 2 (FNC), following the same conservative principle applied to unparseable outputs.

Statistical testing. Pairwise differences between prompt formats are assessed with the paired bootstrap test [Koehn 2004] ($B = 10{,}000$ resamples, $\alpha = 0.05$, two-sided), following recommendations for NLP evaluation [Dror et al. 2018, Berg-Kirkpatrick et al. 2012], resampling instance-level prediction pairs from the majority-voted outputs. Comparisons target RQ2 (*Inst-only* vs. *Demos-only*, and *Inst+Demos* vs. each isolated format) and RQ3 (pairwise among combined formats). Beyond significance, differences are read against a practical-equivalence band of ± 0.02 $BAcc$, a margin no larger than the mean ordering-induced $BAcc$ range observed under *Inst+Demos* (2.5 pp, Section 4.2), so that differences inside the band are of the same order as the variation the protocol itself introduces.

3.5. Implementation

SLMs were evaluated in BFloat16 precision with greedy decoding (`do_sample=False`) and `max_new_tokens=10` (20 for JSON-Schema), using Hugging Face pipelines[3] for both generation and classification tasks, while DeepSeek-V3.2 was accessed via its official API with `temperature=0` to approximate deterministic behavior; all

[3] `https://huggingface.co/docs/transformers/en/main_classes/pipelines`

models operated in standard (non-reasoning) modes. Reproducibility was ensured through fixed random seeds for dataset subsampling and demonstration selection ($s = 42$) and three consistent demonstration orderings ($s \in \{42, 2, 3\}$) applied across configurations. The evaluation comprised 337,479 inference calls spanning all models, datasets, prompt formats, and permutations, with local runs on NVIDIA A100 (40 GB) GPUs via Google Colab; code and full prompt templates are provided in the supplementary repository[2].

4. Experimental Results

4.1. RQ1 – ICL performance across model scales

SLMs surpass the fine-tuned encoder and approach the 685B ceiling. Under their best prompt, all three 9B models exceed MSA-DistilBERT on every dataset by +0.12 to +0.20 *BAcc*, covering 71–101% of the weak-to-strong gap, with 9 of 12 model–dataset cells above 85% (Table 4). Qwen3.5 attains the highest mean coverage (94.5%) and is the only 9B model to reach the 685B ceiling on any dataset (ReLI, 101%). The lower bound falls on the smallest test set (Gemma2 on OPC, N=123), where a single instance shifts gap coverage by roughly 5 pp.

Prompt sensitivity persists across scales but varies by model. Sensitivity does not vanish at 685B: DeepSeek's best and worst prompts still differ by 8.5 pp *BAcc* on average, and by 12.6 pp on TSBr. What scale does not predict is its magnitude. Qwen3.5 matches DeepSeek on both dimensions (mean Δ 0.090 vs. 0.085; CV 0.022 vs. 0.019), while Boto and Gemma2, identical in architecture and initialization, differ by 1.8× in format sensitivity and 3.2× in ordering sensitivity, with Boto's extreme case on TSBr ($\Delta = 0.614$) coinciding with 18.1% FNC. The two dimensions rank consistently together across the 16 model–dataset cells (Spearman $\rho = 0.76$, $p < 0.01$). Sensitivity is thus tied less to scale than to the model itself, motivating the analysis of prompt components in RQ2.

4.2. RQ2 – Disentangling instruction and demonstration contributions

Instructions outperform demonstrations in isolation on most datasets. *Inst-only* achieves significantly higher *BAcc* than *Demos-only* in 7 of 12 model–dataset comparisons (Table 5, top; Figure 1a), with a mean $\Delta BAcc$ of +0.122. The effect is largest for Boto (+0.213), driven by a 27.0 pp FNC reduction, and smallest for Qwen3.5 (+0.040), whose negligible FNC difference (+0.6 pp) indicates a gain in classification rather than in format adherence; Gemma2 lies in between (12.9 pp). ReLI (72.7% Neutral) is the consistent exception: *Demos-only* numerically outperforms *Inst-only* for all three models (significantly for Gemma2 and Qwen3.5), and adding instructions to demonstrations yields no gain.

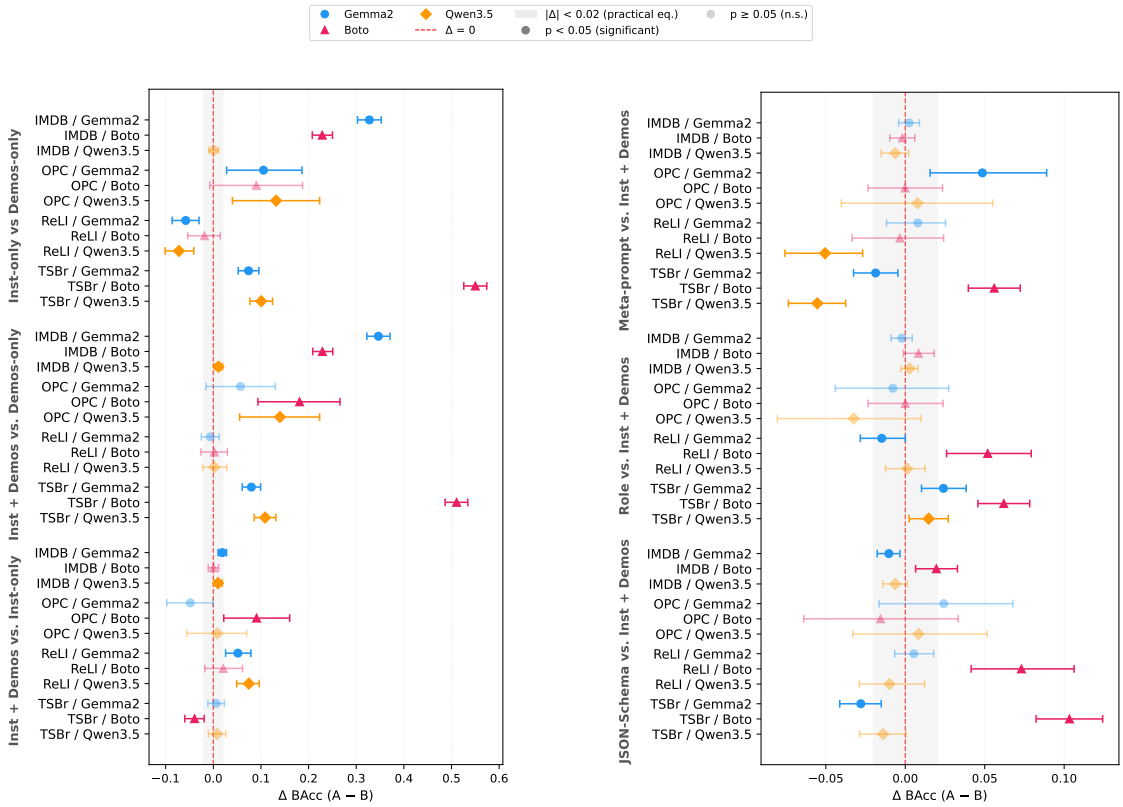
Demonstrations add limited value once instructions are present. *Inst+Demos* significantly outperforms *Demos-only* (mean $\Delta = +0.138$, 8 ↑), but the advantage over *Inst-only* is only +0.017 (5 ↑ 1 ↓), with virtually unchanged FNC rates (+0.3 pp). The forest plot (Figure 1a) confirms that most confidence intervals are narrow and

Table 4. Performance and sensitivity by model × dataset. *Best BAcc*: highest across the six prompts (majority-voted). *Gap*: percentage of the weak-to-strong gap covered, $(\text{BAcc} - \text{MSA})/(\text{DeepSeek} - \text{MSA}) \times 100$, computed from unrounded values; the two baselines anchor the scale at 0 and 100 by construction. Δ: max–min BAcc across prompts (format sensitivity). *CV*: σ/μ of BAcc across the three orderings, averaged over the five few-shot formats, with the mean BAcc range (Δpp) in parentheses. *FNC %*: mean across formats. Bold/underline: best/second-best per dataset (the Gap ranking is identical by construction).

				Format	Ordering	
Model	Dataset	Best BAcc	Gap %	Δ	CV (Δpp)	FNC %
MSA-DistilBERT	IMDB (BSC)	0.755	0		—	
	OPC (BSC)	0.652			—	
	ReLI (MSC)	0.572			—	
	TSBr (MSC)	0.507			—	
Gemma2	IMDB (BSC)	0.941	95	0.349	0.030 (3.8)	7.5
	OPC (BSC)	0.773	71	0.106	0.040 (5.2)	2.8
	ReLI (MSC)	0.711	79	0.060	0.012 (1.7)	1.3
	TSBr (MSC)	<u>0.710</u>	96	0.104	0.028 (3.4)	0.3
	Mean	—	*85.3*	*0.155*	*0.027 (3.5)*	*3.0*
Boto	IMDB (BSC)	0.938	94	0.249	0.051 (7.6)	2.8
	OPC (BSC)	<u>0.813</u>	95	0.181	0.038 (5.2)	0.4
	ReLI (MSC)	0.730	90	0.094	0.051 (6.4)	2.7
	TSBr (MSC)	0.679	81	0.614	0.208 (6.9)	18.1
	Mean	—	*90.0*	*0.284*	*0.087 (6.6)*	*6.0*
Qwen3.5	IMDB (BSC)	<u>0.948</u>	99	0.013	0.006 (1.1)	0.9
	OPC (BSC)	0.797	86	0.148	0.032 (4.5)	2.3
	ReLI (MSC)	**0.749**	101	0.076	0.019 (2.5)	0.1
	TSBr (MSC)	0.703	92	0.123	0.033 (3.9)	1.4
	Mean	—	*94.5*	*0.090*	*0.022 (3.0)*	*1.2*
DeepSeek	IMDB (BSC)	**0.951**	100	0.029	0.005 (0.9)	0.2
	OPC (BSC)	**0.822**		0.113	0.023 (3.2)	0.5
	ReLI (MSC)	<u>0.747</u>		0.072	0.020 (2.8)	0.4
	TSBr (MSC)	**0.719**		0.126	0.028 (3.5)	0.7
	Mean	—	—	*0.085*	*0.019 (2.6)*	*0.5*

centered near zero. In practice, zero-shot instructions capture most of the combined format's performance on balanced or moderately skewed datasets, whereas demonstration-only prompts risk substantial degradation.

Instructions also stabilize predictions against ordering perturbations. Across the three orderings, the mean *BAcc* range drops from 9.6 pp (CV = 0.122) under *Demos-only* to 2.5 pp (CV = 0.018) under *Inst+Demos*, a 3.8× narrowing of the range that holds for every model, DeepSeek at 685B included (4.4×); the stabilizing role of instructions is therefore not scale-dependent (cf. Section 4.1). Gemma2 on IMDB is the clearest case: *Demos-only* yields 0.522, 0.586 and 0.678 across orderings (15.6 pp), converging to 0.938–0.940 (0.2 pp) once instructions are added. Per-ordering trajectories for all six formats, disaggregated by model × dataset,



(a) Decomposing instruction and demonstration contributions: isolated formats and their combination.

(b) Elaborated vs. minimal combined formats (practical-equivalence band).

Figure 1. Paired Bootstrap Test Results ($B = 10,000$, $\alpha = 0.05$). Predictions consolidated via majority voting across 3 demonstration orderings.

are provided in the supplementary repository[4].

4.3. RQ3 – Elaborated vs. minimal combined formats

No elaborated format consistently outperforms the simplest combined baseline. Six of the nine model × format mean differences fall inside the ±0.02 practical-equivalence band (Table 5, bottom; Figure 1b), and the significant comparisons that do occur alternate in sign: Gemma2 gains and loses in equal measure under *Meta-prompt* and *Roleplay*, and is significantly worse under *JSON-Schema* (2 ↓), while Qwen3.5 shows no net change under *Roleplay* or *JSON-Schema* (−0.004, −0.005) and degrades under *Meta-prompt* (−0.026, 2 ↓). Elaboration is thus not a reliable route to accuracy once a task instruction is present.

Boto is the only model that benefits. Elaboration helps precisely the model with the weakest baseline compliance (Section 4.1): *JSON-Schema* (+0.045, 3 ↑), *Roleplay* (+0.031, 2 ↑) and *Meta-prompt* (+0.013, 1 ↑). All three cut Boto's FNC

[4]Direct link: `https://github.com/AndreSchuck/slm_prompt_analysis/blob/master/Article/Images/rq2_ordering_sensitivity_6fmt.png`

Table 5. Summary of paired bootstrap tests ($B = 10{,}000$, $\alpha = 0.05$): each cell reports the mean $\Delta BAcc$ and the number of significant comparisons (↑/↓ favoring the first/second system, – none), out of 4 per model and 12 overall. Top: RQ2; bottom: RQ3.

Comparison	Gemma2	Boto	Qwen3.5	All
Inst-only vs. Demos-only	+0.112 (3 ↑ 1 ↓)	+0.213 (2 ↑)	+0.040 (2 ↑ 1 ↓)	+0.122 (7 ↑ 2 ↓)
Inst+Demos vs. Demos-only	+0.119 (2 ↑)	+0.231 (3 ↑)	+0.065 (3 ↑)	+0.138 (8 ↑)
Inst+Demos vs. Inst-only	+0.007 (2 ↑)	+0.018 (1 ↑ 1 ↓)	+0.025 (2 ↑)	+0.017 (5 ↑ 1 ↓)
Meta vs. Inst+Demos	+0.010 (1 ↑ 1 ↓)	+0.013 (1 ↑)	−0.026 (2 ↓)	−0.001 (2 ↑ 3 ↓)
Role vs. Inst+Demos	−0.000 (1 ↑ 1 ↓)	+0.031 (2 ↑)	−0.004 (1 ↑)	+0.009 (4 ↑ 1 ↓)
JSON vs. Inst+Demos	−0.002 (2 ↓)	+0.045 (3 ↑)	−0.005 (–)	+0.013 (3 ↑ 2 ↓)

rate by a similar margin (≈3.8–4.0 pp) despite differing in whether they constrain output structure, so the benefit tracks instructional specificity rather than format constraints per se.

5. Discussion

The findings across RQ1–RQ3 converge on a coherent narrative: effective ICL-based SA in PT-BR with 9B instruction-tuned SLMs depends less on prompt sophistication than on the presence of a well-formed instruction.

5.1. Scale vs. Performance: 9B SLM Competitiveness

The RQ1 results show that 9B SLMs under ICL can substantially reduce the performance gap to a 685B state-of-the-art model, reinforcing recent findings for PT-BR sentiment classification [Schuck et al. 2025, Lima et al. 2026]. This competitiveness, however, coexists with the known volatility of ICL, where performance depends strongly on prompt design and demonstration ordering [Sclar et al. 2024, Schoch and Ji 2025]. Within this context, results indicate that prompt sensitivity at the 9B scale can match that of 685B models, as Qwen3.5 achieves similar robustness to DeepSeek, possibly reflecting multilingual pre-training and more recent model generation. This suggests that sensitivity is driven more by training regime than scale, implying that, in resource-constrained settings, well-chosen 9B models with effective prompting can recover most of the performance of large proprietary LLMs.

5.2. Instructions as the Primary Driver of ICL Performance

The convergence of RQ2 and RQ3 highlights a hierarchy of prompt components in which instructions dominate over demonstrations, and additional prompt elaboration yields minimal gains. On balanced or moderately skewed datasets, zero-shot *Inst-only* captures most of the achievable performance, while removing instructions (*Demos-only*) significantly harms accuracy and output compliance. This aligns with prior work showing that instruction-tuning enables generalizable instruction-following [Wei et al. 2022], whereas demonstrations mainly convey label space, input distribution, and format [Min et al. 2022]; since these are already encoded in the instructions, their contribution is marginal, explaining the small improvement from *Inst-only* to *Inst+Demos* (mean $\Delta BAcc = +0.017$) and negligible FNC reduction.

5.3. Targeted Fine-tuning and Increased Prompt Sensitivity: A Hypothesis

A counterintuitive result arises in the sensitivity analysis: Boto, the only PT-BR–specific fine-tuned model, shows the highest sensitivity in both format and ordering, whereas Qwen3.5, a natively multilingual model of the same scale, matches the stability of the 685B baseline. One plausible account is that LoRA adaptation to a narrow formatting distribution (*cetacean-ptbr*) weakened the base model's instruction-following behavior and its generalization to unseen prompt structures. This reading rests on a single base→fine-tuned pair, however, and should be taken as a hypothesis rather than a demonstrated causal effect: catastrophic forgetting of instruction-following [Luo et al. 2025], interactions between the fine-tuning chat template and our prompt formats, and LoRA configuration choices [Biderman et al. 2024] are equally consistent with the observed pattern.

6. Conclusion

This work evaluated six prompt formats across three linguistically diverse 9B SLMs and four PT-BR sentiment datasets, anchored by a fine-tuned encoder (0.1B) and a 685B MoE model under a unified protocol. Three findings emerged: (i) 9B SLMs recover 71–101% of the weak-to-strong gap under optimal prompting, with Qwen3.5 matching the 685B baseline in robustness; (ii) instructions, not demonstrations, drive ICL performance on balanced and moderately skewed datasets, with zero-shot *Inst-only* capturing most of the achievable accuracy and instructions narrowing the ordering-induced *BAcc* range by 3.8× (*Demos-only* → *Inst+Demos*), while *Demos-only* degrades both accuracy and output compliance; and (iii) elaborate formats (*Roleplay*, *JSON-Schema*, *Meta-prompt*) provide no consistent gain over *Inst+Demos*. The PT-BR fine-tuned Boto showed the highest sensitivity, suggesting that narrow-distribution LoRA adaptation may reduce robustness. Future work should extend this analysis to other NLP tasks (e.g., QA, NER, NLI), broader scales (1–3B, 20–70B), and additional languages, and examine the interaction between demonstration selection and prompt format in highly skewed settings such as ReLI.

Limitations

This study has several limitations that inform result interpretation and future work. The evaluation is restricted to SA, the 9B parameter range, and four PT-BR datasets, limiting generalization across tasks, scales, and languages [Peres 2023, Alves et al. 2023, Geng et al. 2024]; exploring smaller (1–3B) and larger (20–70B) models could clarify whether prompt sensitivity decreases with scale. The cross-scale comparison in RQ1 is exploratory, as models differ in architecture (MoE vs. dense), execution (API vs. local BFloat16), and reproducibility, so reported differences in amplitude and CV are indicative rather than causal. The high accuracy on IMDB-PT may reflect cross-lingual contamination, since it derives from a widely used English benchmark [Maas et al. 2011].

Methodological trade-offs include discarding Neutral probabilities in MSA-DistilBERT, a conservative bias in majority-voting ties toward FNC, and a fixed `max_new_tokens` that can penalize verbose formats. Few-shot results also rest on a single demonstration-selection strategy (k-means over *jina-embeddings-v3*), which maximizes semantic diversity but presumes access to a labeled training split and to an embedding model, a non-trivial requirement in production settings. Since selection is known to interact with ordering sensitivity [Liu et al. 2021a, Lu et al. 2022], and RQ2 indicates that demonstrations contribute marginally once instructions are present, we expect this choice to have limited impact here, although a random-selection control remains necessary to confirm it.

Ethics Statement

This work uses only publicly available datasets (IMDB-PT, OPCovid-BR, ReLI, TweetSentBr) and pre-trained models accessed under their respective licenses. No human-subject data was collected. Automated SA systems may amplify linguistic biases in training data when deployed in high-stakes contexts; our findings are presented strictly as a contribution to the scientific understanding of prompt engineering for PT-BR NLP.

Generative AI Usage

Generative AI was used in two capacities: Claude 3.5 Sonnet (Anthropic) produced the instruction component of the Meta-prompt format (one of six evaluated configurations); LLM tools also assisted research, manuscript drafting and revision. All scientific claims and conclusions are the sole responsibility of the authors.

Acknowledgments

This study was partially funded by the São Paulo Research Foundation (FAPESP), Brazil, under Grant Nos. #2025/13172-1, #2024/22853-0, and #2023/14427-8. This work was also supported by Petrobras through research grants No. #2025/00607-0 and No. #2023/00466-1.

References

Ajith, A., Pan, C., Xia, M., et al. (2023). Instructeval: Systematic evaluation of instruction selection methods. https://arxiv.org/abs/2307.00259.

Alves, D. M., Guerreiro, N. M., Alves, J., et al. (2023). Steering large language models for machine translation with finetuning and in-context learning. https://arxiv.org/abs/2310.13448.

Assis, G., Amorim, A., Carvalho, J., et al. (2024). Exploring Portuguese hate speech detection in low-resource settings: Lightly tuning encoder models or in-context learning of large models? In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 301–311, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Berg-Kirkpatrick, T., Burkett, D., and Klein, D. (2012). An empirical investigation of statistical significance in NLP. In Tsujii, J., Henderson, J., and Paşca, M., editors, *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea. Association for Computational Linguistics.

Biderman, D., Portes, J., Ortiz, J. J. G., Paul, M., Greengard, P., Jennings, C., King, D., Havens, S., Chiley, V., Frankle, J., Blakeney, C., and Cunningham, J. P. (2024). Lora learns less and forgets less.

Birjali, M., Kasri, M., and Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, 226:107134.

Brígida, L. S. (2024). Boto 9b it. https://huggingface.co/lucianosb/boto-9B-it.

Brown, C. E. (1998). Coefficient of variation. In *Applied multivariate statistics in geohydrology and related sciences*, pages 155–157. Springer.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners.

DeepSeek-AI (2025). Deepseek-v3.2: Pushing the frontier of open large language models.

Devlin, J., Chang, M.-W., Lee, K., et al. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple Classifier Systems*, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg.

Dong, Q., Li, L., Dai, D., et al. (2024). A survey on in-context learning. https://arxiv.org/abs/2301.00234.

Dror, R., Baumer, G., Shlomov, S., et al. (2018). The hitchhiker's guide to testing statistical significance in natural language processing. In Gurevych, I. and Miyao, Y., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Fu, J., Ng, S.-K., and Liu, P. (2022). Polyglot prompt: Multilingual multitask promptraining. https://arxiv.org/abs/2204.14264.

Gao, S., Wen, X.-C., Gao, C., et al. (2023). What makes good in-context demonstrations for code intelligence tasks with llms? In *2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE.

Gemma Team, Riviere, M., Pathak, S., et al. (2024). Gemma 2: Improving open language models at a practical size. https://doi.org/10.48550/arXiv.2408.00118.

Geng, M., Wang, S., Dong, D., et al. (2024). Large language models are few-shot summarizers: Multi-intent comment generation via in-context learning. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, ICSE '24, New York, NY, USA. Association for Computing Machinery.

Grandini, M., Bagli, E., and Visani, G. (2020). Metrics for multi-class classification: an overview. https://arxiv.org/abs/2008.05756.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In Lin, D. and Wu, D., editors, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Lima, J. V. R. J., Pinheiro, V., and Caminha, C. (2026). Portuguese sentiment analysis with open-source LLMs: Models, prompts, and efficient deployment. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 212–221, Salvador, Brazil. Association for Computational Linguistics.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Springer International Publishing.

Liu, J., Shen, D., Zhang, Y., et al. (2021a). What makes good in-context examples for gpt-3? https://arxiv.org/abs/2101.06804.

Liu, P., Yuan, W., Fu, J., et al. (2021b). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. https://doi.org/10.48550/arXiv.2107.13586.

Lu, Y., Bartolo, M., Moore, A., et al. (2022). Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. https://arxiv.org/abs/2104.08786.

Lu, Z., Li, X., Cai, D., et al. (2025). Small language models: Survey, measurements, and insights. https://arxiv.org/abs/2409.15790.

Luo, Y., Yang, Z., Meng, F., Li, Y., Zhou, J., and Zhang, Y. (2025). An empirical study of catastrophic forgetting in large language models during continual fine-tuning.

Maas, A. L., Daly, R. E., Pham, P. T., et al. (2011). Learning word vectors for sentiment analysis. https://aclanthology.org/P11-1015.

Marreira, E., De Melo, T., De Oliveira, M., et al. (2025). Rating prediction in brazilian portuguese: A benchmark of large language models. *Journal of the Brazilian Computer Society*, 31:828–839.

Min, S., Lyu, X., Holtzman, A., et al. (2022). Rethinking the role of demonstrations: What makes in-context learning work? https://arxiv.org/abs/2202.12837.

Nguyen, C. V., Shen, X., Aponte, R., et al. (2025). A survey on small language models. In Angelova, G., Kunilovskaya, M., Escribe, M., and Mitkov, R., editors, *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 807–821, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Oliveira, A., Nascimento, E., Pinheiro, J., et al. (2025). Small, medium, and large language models for text-to-sql. In Maass, W., Han, H., Yasar, H., and Multari, N., editors, *Conceptual Modeling*, pages 276–294, Cham. Springer Nature Switzerland.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2:1–135.

Pereira, D. A. (2021). A survey of sentiment analysis in the portuguese language. *Artif. Intell. Rev.*, 54(2):1087–1115.

Peres, R. S. (2023). Grandes modelos de linguagem na resolução de questões de vestibular: o caso dos institutos militares brasileiros. Master's thesis, Universidade Federal do Estado do Rio de Janeiro.

Piorino, G., Moreira, V., Lima, L. H. Q., et al. (2025). Sentiment analysis of shared content in brazilian reddit communities. *Journal on Interactive Systems*, 16:666–686.

Qwen Team (2026). Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5.

Reynolds, L. and McDonell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. https://arxiv.org/abs/2102.07350.

Sanh, V., Debut, L., Chaumond, J., et al. (2020). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. https://arxiv.org/abs/1910.01108.

Schoch, S. and Ji, Y. (2025). The good, the bad, and the debatable: A survey on the impacts of data for in-context learning. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V., editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29798–29812, Suzhou, China. Association for Computational Linguistics.

Schuck, A. d. F., Garcia, G. L., Manesco, J. R. R., et al. (2025). Evaluating large language models for brazilian portuguese sentiment analysis: A comparative study

of multilingual state-of-the-art vs. brazilian portuguese fine-tuned llms. *Journal of the Brazilian Computer Society*, 31(1):885–917.

Schulhoff, S., Ilie, M., Balepur, N., et al. (2024). The prompt report: A systematic survey of prompting techniques. https://arxiv.org/abs/2406.06608.

Sclar, M., Choi, Y., Tsvetkov, Y., and Suhr, A. (2024). Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting.

Shin, T., Razeghi, Y., au2, R. L. L. I., et al. (2020). Autoprompt: Eliciting knowledge from language models with automatically generated prompts. https://arxiv.org/abs/2010.15980.

Simmering, P. F. and Huoviala, P. (2023). Large language models for aspect-based sentiment analysis. https://arxiv.org/abs/2310.18025.

Sturua, S., Mohr, I., Akram, M. K., et al. (2024). jina-embeddings-v3: Multilingual embeddings with task lora. https://arxiv.org/abs/2409.10173.

TabularisAI, Gyamfi, S., Borisov, V., et al. (2025). multilingual-sentiment-analysis (revision 69afb83). https://huggingface.co/tabularisai/multilingual-sentiment-analysis.

Wang, F., Lin, M., Ma, Y., et al. (2025). A survey on small language models in the era of large language models: Architecture, capabilities, and trustworthiness. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, pages 6173–6183, New York, NY, USA. Association for Computing Machinery.

Wang, X., Wei, J., Schuurmans, D., et al. (2023). Self-consistency improves chain of thought reasoning in language models. https://arxiv.org/abs/2203.11171.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. (2022). Finetuned language models are zero-shot learners.

Wei, J., Wei, J., Tay, Y., et al. (2023). Larger language models do in-context learning differently. https://arxiv.org/abs/2303.03846.

White, J., Fu, Q., Hays, S., et al. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. https://arxiv.org/abs/2302.11382.

Xu, H., Wang, Q., Zhang, Y., Yang, M., Zeng, X., Qin, B., and Xu, R. (2024). Improving in-context learning with prediction feedback for sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3879–3890.

Zhao, W. X., Zhou, K., Li, J., et al. (2024). A survey of large language models. https://arxiv.org/abs/2303.18223.

Zheng, M., Pei, J., Logeswaran, L., et al. (2024). When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. https://arxiv.org/abs/2311.10054.

Zhou, Y., Muresanu, A. I., Han, Z., et al. (2023). Large language models are human-level prompt engineers. https://arxiv.org/abs/2211.01910.

Compression-Based Linguistic Complexity Metrics in Automatic Essay Scoring

Felipe Ribas Serras[1], **Igor Cataneo Silveira**[1], **Denis Deratani Mauá**[1], **Marcelo Finger**[1]

[1]Department of Computer Science
Institute of Mathematics, Statistics and Computer Science (IME)
University of São Paulo (USP) – São Paulo, SP – Brazil

{frserras, igorcs, ddm, mfinger}@ime.usp.br

Abstract. *Compression-based linguistic complexity metrics enable cross-linguistic comparison without prior annotation. Their sensitivity to variation across languages and Portuguese registers highlights their applicability in NLP tasks. This study investigates their use as readability proxies and complementary features in Automatic Essay Scoring. We analyze how these metrics capture variation in essay quality across traits, genres, and educational levels in Brazilian Portuguese. In addition, we evaluate their sensitivity to differences between human- and AI-generated essays. Our results suggest that complexity metrics are effective (i) in differentiating educational levels, (ii) in detecting whether they were written by humans and (iii) as predictors of essay quality.*

1. Introduction

Variation in structural linguistic complexity, the way in which linguistic manifestations employ strategies to distribute information across different linguistic subsystems, is a broad linguistic phenomenon that spans both the typological level, in terms of structural variation between languages [Croft 2002], and the sociolinguistic level, regarding how a single language adapts to different social contexts [Biber and Conrad 2009].

Structural complexity is an important parameter in various linguistic hypotheses and models, such as the equi-complexity hypothesis, the morphosyntactic trade-off hypothesis [Hockett 1958], and Multidimensional Analysis models of intralinguistic variation [Biber 1995]. It is also a major feature in the development and understanding of language processing technologies, such as linguistic simplification and elaboration models [Leal et al. 2026].

Grounded in information theory, compression-based linguistic complexity metrics [Juola 1998, Juola 2008, Ehret and Szmrecsanyi 2016] provide an automated approach to estimating structural linguistic complexity. By using off-the-shelf data compressors and random text perturbation, they measure the information conveyed at specific linguistic subsystems, such as syntax and morphology.

Because they operate directly on raw text rather than relying on annotation frameworks, these metrics are holistic and function comparably across diverse languages and linguistic scopes. Consequently, they offer a highly robust tool for analyzing transversal complexity phenomena, capturing evidence that emerges simultaneously across different scopes of linguistic variation, both intralinguistic and cross-linguistic. These metrics have demonstrated their descriptive power and sensitivity to different forms of variation, including typological variation across different languages and linguistic groups

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 390–404
October 19-22, 2026 ©2026 Association for Computational Linguistics

[Juola 2008, Serras et al. 2024], and diachronic variation across historical variants of the same language [Ehret and Szmrecsanyi 2016].

In particular, this family of metrics has proven especially sensitive to register variation in both English [Ehret 2021] and Portuguese [Serras and Finger 2026]. Registers are varieties within a single language that emerge as a functional response to the context of use. Differences in register may vary in intensity, but they are pervasive throughout a text and are socially shared and mediated by a linguistic community, such as informal conversation, online forum discussions, and legal petitions [Biber and Conrad 2009].

One context in which the pervasive adaptation of language to a target register is instrumental is that of essay assessments, where candidates seek to demonstrate linguistic mastery by adhering to that register. Given the sensitivity of compression-based metrics to register variation, it is worth investigating how sensitive these metrics are to grade variation within this context, and whether they could serve as interpretable or supplementary features in automated essay scoring systems.

Automated Essay Scoring (AES) is an NLP task that aims to assign grades within a pre-established trait framework based on an essay's content [Page 1966]. These systems can serve as (i) support tools for graders, (ii) quality control mechanisms to ensure consistency in grading for institutions, and (iii) feedback proxies for student self-practice, especially for those with limited access to formal educational resources. In Brazilian Portuguese, these systems focus specially on essays required for the ENEM [Amorim and Veloso 2017, Marinho et al. 2021, Silveira et al. 2024, Rossman et al. 2026], the Brazilian National High School Exam, a large-scale and high-stakes exam that serves as the national standardized test for university admission, which requires the production of an argumentative essay on a different prompt every year.

The advent of Transformer-based Large Language Models has brought a significant increase in performance to AES systems [Boquio and Naval 2024]. In general, but especially for Portuguese, the state-of-the-art results in AES are currently those obtained by this family of models [Mello et al. 2024, Barbosa et al. 2025]. In this context, explainability is one of the greatest challenges for those systems, which function as black boxes. This makes it difficult to understand both the means by which the models achieve this performance and the rationale behind the scores they assign.

Readability metrics have being widely used in the context of AES for Portuguese, as interpretable features for models [Marinho et al. 2022, Silveira et al. 2025b, Chalegre et al. 2026] and also as means of distinguishing different essay sub-registers [Cavalcanti et al. 2025]. However, the inclusion of new features in these analyzes can consistently offer a fresh perspective on how AES models operate and the differences and peculiarities of the diverse range of essay sub-registers to which they are applicable.

To this end, we seek to integrate these two areas of investigation in this article by examining the sensitivity of compression-based linguistic complexity metrics to grade variation in AES datasets across different educational levels and essay sub-registers. By doing so, we intend to better understand the sensitivity and robustness of these metrics, generating a deeper understanding of their behavior in Portuguese and ascertaining their usability as AES features for both model training and analysis.

Inspired by [Serras and Finger 2026], we conducted a statistical analysis of the

sensitivity of these metrics, both to scores and to differences between sub-registers of essays, using three AES datasets in Brazilian Portuguese: the Brazilian Portuguese Narrative Essays Dataset [Mello et al. 2024], AES-ENEM [Silveira et al. 2024], and Diplomatrix-BR [Cavalcanti et al. 2025]. These datasets cover narrative and argumentative essays produced by children, adolescents, adults, and language models. Through this study, we aimed to answer the following research questions:

1. How well do compression-based complexity metrics predict the scores assigned to essays across different sub-registers, educational levels, and traits?
2. How well do these metrics distinguish among different essay sub-registers?
3. How well do these metrics distinguish between human-written and LLM-generated essays?
4. What do these results reveal about structural linguistic variation across essays and the sensitivity of this specific family of metrics?
5. Could these metrics be incorporated into AES systems as features, or aid in their interpretation?

Overall, we conclude that these metrics are quite effective at distinguishing among different essay sub-registers and between human-written and synthetic essays. In many cases, they serve as statistically significant predictors of the assigned scores, particularly where there is greater structural variation and where adherence to linguistic structure is assessed more directly, such as in essays written by children who are still mastering these structures. Nevertheless, they generally act as weak to moderate score predictors; while they can be valuable tools for interpreting these models, they should be used in conjunction with other metrics to provide a more accurate picture of their behavior.

2. Related Work

In this section, we review related work, categorizing it into the two main areas of research covered by this study: linguistic complexity metrics and AES.

2.1. Linguistic Complexity Metrics

The modern approach to structural linguistic complexity as a parameter stems from Hockett's hypotheses [Hockett 1958]. [Nichols 1998] and [McWhorter 2001] are examples of the earliest proposals to characterize this form of complexity as a quantitative parameter, involving subjective interpretation and manual annotation by experts.

Methods for automatically computing complexity first emerged through readability metrics in the field of language acquisition [Leal et al. 2026]. Tools such as Coh-Metrix [McNamara et al. 2014] for English, and NILC-Metrix [Leal et al. 2024] and iRead4Skills [Monteiro et al. 2025, Aissa et al. 2025] for Portuguese represent the culmination of this approach, enabling the automated extraction of various complexity and readability features from text. Although powerful, these tools rely on annotation systems that vary across languages, limiting their applicability in cross-linguistic parameterization.

Compression-based Complexity metrics have emerged as alternative text-based approaches that are grounded in information theory and language-agnostic, enabling the comparability of results across different languages and linguistic scopes.

Proposed in [Juola 1998, Juola 2008] and refined in [Ehret and Szmrecsanyi 2016], they were validated against cross-linguistic [Juola 2008, Serras et al. 2024], diachronic [Ehret and Szmrecsanyi 2016], and synchronic intralinguistic variation in English [Ehret 2021] and Portuguese [Serras and Finger 2026]. The latter studies explored variation at the register level, where broader linguistic descriptions from Multidimensional Analysis Models point to a strong dependence on structural complexity and informational density across various languages [Biber and Conrad 2009, Biber 1995], including Portuguese [Sardinha et al. 2014, Serras et al. 2026].

More recently, readability and complexity features have also been explored in the characterization and detection of synthetic texts generated by LLMs. [Lobo and Martins 2026], for example, examines the differences between Portuguese scientific abstracts with and without LLM influence, using NILC-Metrix features. [Cavalcanti et al. 2025] presents Diplomatrix-BR, an essay dataset comprising both human-authored texts and essays generated by various LLMs in response to the same prompts; furthermore, using NILC-Metrix features, the authors analyze the differences between these groups. Readability features were also used by [Locatelli et al. 2024] and [Liu et al. 2023] to compare synthetic and natural argumentative essays in Portuguese and English, respectively.

2.2. Automatic Essay Scoring

Although early AES systems relied on a feature-based approach combined with linear regression [Page 1966], subsequent NLP technologies improved model generalization at the expense of interpretability. However, in this domain, explicability is critical, and it is often worthwhile to sacrifice predictive performance in favor of interpretability. For this reason, modern approaches seek to balance both objectives.

For Portuguese, several sets of interpretable features have been used to this end. Examples are: the words in the essay [Bazelato and Amorim 2013]; a combination of domain-relevant features with features adapted from English [Amorim and Veloso 2017]; n-grams, and POS-tagging n-grams [Fonseca et al. 2018]; and ENEM-trait-specific feature sets [Marinho et al. 2022]. Recently, NILC-Metrix has been widely employed as it offers a comprehensive suite of features within a single framework. It has been used (i) to calculate proxies for what Transformer-based models may be learning [Silveira et al. 2025a], (ii) as baseline performance when combined with a linear regressor and random forest [Barbosa et al. 2025], and (iii) as an alternative input representation for neural networks [Chalegre et al. 2026].

Other works aim to combine Transformers and feature explainability. One approach combines them by adding the Transformer-assigned score to a pool of interpretable features [Liu et al. 2019, Silveira et al. 2025b], or by concatenating the features with dense representations [Matos and Feltrim 2026]. Alternatively, another approach employs Transformers to predict specific essay characteristics, subsequently combining them through (interpretable) logic [Silveira and Mauá 2026].

3. Methodology

In this section, we describe our complexity metrics (Section 3.1), the AES datasets that we examined (Section 3.2), the statistical methods used to assert the correlation between

the metrics and the grades in each dataset (Section 3.3), and the methods used to attest the sensitivity of the metrics to differentiate between essay sub-registers (Section 3.4). The code, plots, and results from all described procedures are available in our repository[1].

3.1. Compression-based Complexity Metrics

Compression-based complexity metrics reduce language to its communicative function of transmitting information. In this view, the complexity of a text T is defined as the amount of information it conveys or the shortest representation of the text from which it can be reconstructed, known as Kolmogorov complexity. The Kolmogorov complexity of T is generally uncomputable but can be approximated using data compression algorithms, such as `gzip`, by generating a compressed version $K(T)$ of the original text and measuring its size in bits, represented by $|K(T)|$ [Grünwald and Vitányi 2008].

To assess the amount of information conveyed by a specific linguistic subsystem μ, such as syntax s or morphology m, these metrics apply random perturbations to the text, designed to primarily affect the patterns of the target subsystem ($T \rightarrow T^{\mu}$), and measure the effect on the amount of information transmitted ($|K(T^{\mu})|$). The syntactic complexity C_s and morphological complexity C_m metrics are obtained by comparing this perturbed value to the amount of information in the original text, following Equations 1 and 2. In the version of these metrics used in this study, the perturbation applied to the texts consists of randomly deleting 10% of the units within the target subsystem: characters for morphology and entire words for syntax.

$$C_s(T) = \frac{|K(T^s)|}{|K(T)|} \quad (1) \qquad C_m(T) = -\frac{|K(T^m)|}{|K(T)|} \quad (2) \qquad C^*_m(T) = \frac{|K(T)|}{|K(T^{m*})|} \quad (3)$$

Conceptually, C_s measures the importance, in terms of information transmission, of a specific word appearing in a specific position in the text; it therefore estimates the obligatory level of syntactic elements: the more restricted and closed the structure is regarding the presence of its elements, the higher C_s will be. C_m, on the other hand, is a measure of lexical-morphological diversity at the informational level. The more diverse the lexicon and the lexical construction mechanisms in the sample, the higher we expect C_m to be.

An alternative measure of morphological complexity is C^*_m, presented in [Juola 1998]. For this measure, the perturbation applied to morphology is more significant, replacing all words with unique numerical indices. Thus, the perturbed version of the text, T^{m*}, is stripped of all information about the content of the words themselves, retaining only information about their positions. The resulting C^*_m complexity, computed using Equation 3, measures the amount of raw information contained in word representations. In this sense, it is a much coarser measure than C_m, as it seeks to capture the overall morphological information, rather than the subtleties of the sample's morphological system.

For our experiments, we used [Serras et al. 2024]'s implementation of these metrics[2], which computes C_s, C_m, and C^*_m in Python using two distinct compressors: `gzip`, a general-purpose data compressor, and `bzip2`, a compressor optimized for text data.

[1] `https://github.com/frserras/essay-complexity-pt`

[2] `https://github.com/frserras/lang_complexity` (`version 1.0.0`)

3.2. AES Datasets

In this study, we used three datasets:

- The **Brazilian Portuguese Narrative Essays Dataset** [Mello et al. 2024] which comprises 1233 narrative essays written by students from 5th to 9th grade in Brazil on two different prompts. These essays are graded in an ordinal scale containing five classes, according to four different traits: `cohesion`, `formal register`, `narrative rhetorical structure` and `thematic coherence`.
- The **AES-ENEM Dataset** is composed by argumentative essays written for mock ENEM exams [Silveira et al. 2024]. Here essays are also graded in an ordinal scale containing six classes, in five different traits: `(C1) proper usage of Portuguese`, `(C2) thematic coherence`, `(C3) argumentation`, `(C4) coherence` and `(C5) conclusion strength`. Importantly, we removed the lowest class, as it does not reflect text quality but rather the author not adhering to the prompt. This dataset includes grades by two different graders and the original web scrapped grades.
- The **Diplomatrix-BR dataset** [Cavalcanti et al. 2025] is made of argumentative essays written for diplomat civil service tests. This last dataset differs from the previous ones in two ways: a) it is scored on a continuous scale; b) it contains both synthetic and human-written essays.

3.3. Grade-Complexity Correlation Analysis

To assess the predictive power of complexity metrics for essay scores, we computed each metric variant listed in Section 3.1 ten times per essay and averaged the results. We then measured the correlation between each metric and the scores for each trait within each dataset separately.

We chose to categorize the Diplomatrix continuous scores into quintiles, transforming them into an ordinal categorical variable, consistent with the scores in the other two datasets. This allowed us to apply the same correlation test across all datasets. We used Kendall's rank correlation test [Hollander et al. 2014], specifically designed for ordinal categorical variables[Brideau 2025, Laerd Statistics 2018]. Of the three correlation coefficients frequently used for this test, we selected $\tau - b$, which is recommended for scenarios with ties, such as ours [Kendall 1945].

The test is non-parametric [Hollander et al. 2014] and relies on only two assumptions: (i) both variables must be ordinal, which holds true for both complexity metrics and scores; and (ii) there must be an approximately monotonic relationship between the target variables. This latter assumption is not strict; however, when the relationship deviates from monotonicity, there is an increased risk of false negatives [Laerd Statistics 2018]. We examined the relationships across all metric-score pairs: most were approximately monotonic, but we isolated the non-monotonic pairs for a more cautious interpretation.

Since multiple tests were performed simultaneously, we applied the Benjamini-Hochberg False Discovery Rate correction to the p-values to reduce the risk of false positives resulting from multiple testing [Benjamini and Hochberg 1995]. For both the test and the p-value correction, we used the implementations from the `scipy`[3] package.

[3]https://scipy.org/

3.4. Essay Sub-register Sensitivity Analysis

To assess the ability of the metrics to differentiate the essay sub-registers, we used the analysis methodology outlined in [Serras and Finger 2026]: (i) we segmented the essays from all datasets into sentences using the Portuguese sentence tokenizer from the spaCy package[4]; (ii) we randomly selected 2200 sentences from each sub-register to compile the sentence dataset; (iii) we sampled from this dataset 250 sentences per sub-register, combining them in random order to form a pseudo-document, repeating this process 250 times for each sub-register to generate 250 pseudo-documents; (iv) we computed the complexity metrics for each pseudo-document; and (v) we used these values to perform an Analysis of Variance (ANOVA) test to determine if the complexity varies significantly across the sub-registers. In contrast to [Serras and Finger 2026], we employed Welch's ANOVA[5], which does not assume homogeneity of variances across groups, as our exploratory analyses revealed that the sub-registers exhibit markedly distinct width distributions across several metrics [Welch 1951, Liu 2015].

The test computes the F-statistic, which evaluates the ratio of between-group variability to within-group variability. A higher F-value indicates greater separability of the sub-registers for a given complexity metric. Because the F-statistic can be sensitive to sample size, we also calculated the effect size η^2. An η^2 value approaching 1 indicates highly separable sub-registers.

The assumptions of Welch's ANOVA are: (i) independence of observations, which was ensured by the experimental design, in which each data point corresponds to a pseudo-document with an approximate 10% sentence overlap; and (ii) normality of residuals, which we verified by analyzing Quantile-Quantile (Q-Q) plots comparing the observed residuals for each metric against a theoretical normal distribution [Welch 1951, Liu 2015].

We also investigated the morphosyntactic trade-off across essay sub-registers by (i) examining scatter plots of complexity metrics and (ii) assessing the correlation coefficients between these metrics. The morphosyntactic trade-off is a structural pattern wherein higher syntactic complexity tends to co-occur with lower morphological complexity, and vice versa. Compression-based metrics are particularly sensitive to this phenomenon, both at the typological level [Juola 2008, Serras et al. 2024] and across broader linguistic registers in Portuguese [Serras and Finger 2026]. This study allows us to verify whether this pattern persists at this more granular level of variation.

4. Results

The results are divided into two parts: Grade-Complexity Correlation Analysis (Section 4.1) and Essay Sub-register Sensitivity Analysis (Section 4.2).

4.1. Grade-Complexity Correlation Analysis

Figures 1a and 1b present the correlation matrices between the metrics and the score traits for the Narrative dataset, and the AES-ENEM Dataset[6] respectively. Cells marked

[4] `https://spacy.io/`

[5] We utilized the `statsmodels` library (`https://www.statsmodels.org/stable/index.html`)

[6] For brevity, we present results for "Grader A". Full results are in our repository

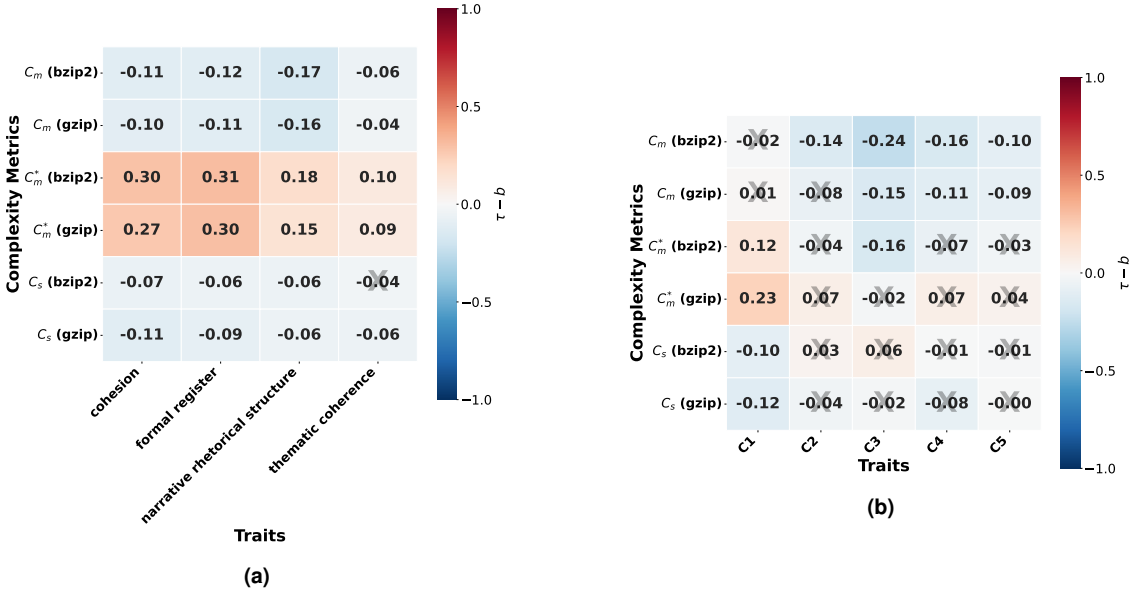


Figure 1. $\tau - b$ **correlation coefficients for (a) the Narrative Dataset and (b) the AES-ENEM Dataset. Cells marked with an X have non-significant associated p-values.**

with 'X" indicate statistically non-significant p-values and are therefore excluded from consideration. In several cases, we observed statistically significant correlation coefficients, indicating that the complexity metrics studied are meaningful predictors of scores. However, these correlations indicate low to moderate predictability, suggesting that these metrics should be used supplementarily rather than independently.

The higher the educational level of the essays, the weaker and less statistically significant the correlations tend to be. Notably, the metrics are not statistically significant predictors for differentiating the quintiles of the Diplomatrix scores, whereas they are significant predictors for all traits of the narrative dataset. These results can be explained by two factors: (i) at more advanced educational levels, candidates are more likely to demonstrate mastery of the desired linguistic structures, producing more cohesive essays with less structural variation; and (ii) at less advanced levels, structural adherence becomes more critical for grading due to substantial variation among candidates. Consequently, although adherence to register is an evaluative criterion at higher educational levels, it is seldom a differentiating factor because the vast majority of candidates achieve high adherence. Conversely, the narrative dataset exhibits notable variation; a qualitative sampling reveals some narrative texts that adhere closely to the expected register, while others present clearly non-adherent structures, especially polysyndetic sentences characteristic of children's writing.

Among all metrics, C_m^*, which represents raw morphological complexity, is the strongest predictor of grades. This metric correlates with text length and word length, indicating that, as expected, longer texts or those with greater lexical complexity or diversity tend to achieve higher evaluation scores. The traits most strongly correlated with this metric are `formal register` and `cohesion` in the Narrative Dataset, and `(C1) proper usage of Portuguese` in the AES-ENEM dataset. The studied structural complexity metrics, known to be sensitive to register variation, align precisely with the

traits assessing adherence to formal register and structure, as expected.

The other metrics exhibit weaker or inverse correlations with the scores, indicating that, at least within these samples, these subtle dimensions of complexity are negative predictors and less prominent than raw morphological complexity. We hypothesize that the direction of these correlations stems from C_m capturing off-topic lexico-morphological diversity and C_s reflecting excessive syntactic rigidity, which negatively impacts the scores. Remarkably, the part of the ENEM dataset that was graded by experts demonstrated higher inter-annotator agreement in terms of correlation with the complexity metrics than the part that was web-scraped, underscoring the consistency and quality of those annotations.

4.2. Essay Sub-register Sensitivity Analysis

Table 1 presents the ANOVA results evaluating the metrics' ability to differentiate writing sub-registers. The large F-statistics demonstrate the strong capacity of these metrics to differentiate registers, even at this higher level of specificity. The large η^2 values reinforce these findings, indicating that they are not an artifact of sample size. Overall, the discriminatory power of the morphological metrics C_m and C_m^* is greater than that of the syntactic metrics C_s, consistent with the observations of [Serras and Finger 2026].

Metric	F	P-Value	η^2
C_m (bzip2)	11344	< 0.001	0.98
C_m (gzip)	11376	< 0.001	0.98
C_m^* (gzip)	10658	< 0.001	0.97
C_m^* (bzip2)	3190	< 0.001	0.93
C_s (gzip)	1334	< 0.001	0.81
C_s (bzip2)	805	< 0.001	0.73

Table 1. ANOVA results between all essay sub-registers

Metric	F	P-Value	η^2
C_m (gzip)	7193	< 0.001	0.94
C_m (bzip2)	3766	< 0.001	0.88
C_m^* (gzip)	2951	< 0.001	0.86
C_s (gzip)	2167	< 0.001	0.81
C_s (bzip2)	1042	< 0.001	0.68
C_m^* (bzip2)	646	< 0.001	0.56

Table 2. ANOVA results between human-authored and LLM-generated essays

To distinguish between LLM-generated and human-authored texts, we conducted a separate ANOVA on the Diplomatrix data. The results, presented in Table 2, further demonstrate the robust capacity of the studied metrics to differentiate these two categories.

When evaluating the trade-off between syntactic and morphological complexity metrics, we excluded the narrative dataset. This subset exhibited significantly different complexity values, likely reflecting the characteristics of children's writing, which allowed for greater sensitivity in the previous analysis. Figure 2 illustrates the trade-off between the remaining sub-registers using the `gzip`-based metric variants. Despite the limited number of data points, the overall trade-off remains evident. Specifically, Figure 2b demonstrates a clear distinction between human-authored essays and those produced by LLMs.

The results suggest that LLM-generated texts are shorter, exhibit reduced lexical variety, and employ more formulaic and rigid syntactic structures than human-authored texts, consistent with the observations of [Liu et al. 2023, Locatelli et al. 2024]. This aligns with the hyper-structuring of these synthetic essays, which often organize the con-

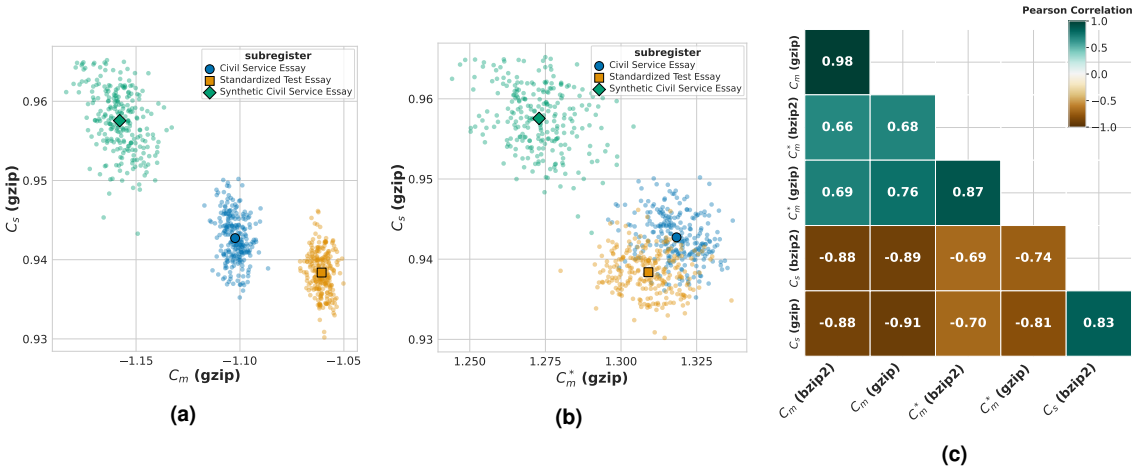


Figure 2. Trade-off plots between Syntactic and Morphological Complexity Metrics using `gzip` (a, b) and Correlation matrix between complexity metrics across essay sub-registers (c).

tent into sections and copy entire parts of the reference text, a behavior rarely observed in human writers.

Conversely, the ENEM Dataset essays exhibit less rigid syntax and greater lexico-morphological diversity than those produced for the diplomatic exam, in accordance with the observations in [Cavalcanti et al. 2025] about diversity and ambiguity. This likely reflects the broader thematic scope of ENEM essays, which cover more comprehensive topics; furthermore, their wider variance in register adherence results in lower syntactic complexity.

The correlations among the various metrics across the sub-registers, illustrated in Figure 2c, support these results, demonstrating an inverse correlation between syntactic and morphological metrics, positive correlations among metrics within the same subsystem, and strongly positive correlations between variants of the same metric using different compressors. This aligns with previous findings suggesting that, at least when using `gzip` and `bzip2`, the effect of variation in data complexity outweighs the variation introduced by the chosen compressor.

5. Conclusions

In this study, we evaluated the sensitivity of compression-based structural complexity metrics to variation in essay sub-registers, as well as their predictive power for scores across different educational levels and traits. Overall, the metrics serve as statistically significant predictors of grades, especially when adherence to the target register structure is a key grading criterion. However, their predictive power tends to be weak to moderate, indicating that they should be used as supplementary features. Furthermore, these metrics effectively differentiate between sub-registers as well as between LLM-generated and human-authored texts, exhibiting a morphosyntactic trade-off and an emphasis on morphological differentiation, consistent with previous findings.

Limitations

The limitations of this study and avenues for future work include (i) small sample sizes and confounded variables, such as the overlap between narrative and children's texts; (ii) the need for more diverse, multilingual datasets to leverage language-agnostic metrics; (iii) the potential of combining explainable metrics with qualitative analysis; (iv) unexplored options in the Diplomatrix dataset, such as alternative continuous-to-categorical conversion methods and model-specific segmentation of synthetic essays; (v) sub-optimal essay lengths for compression-based metrics, although exploring longer texts remains a future path; (vi) a scope limitation to robust statistical exploration without actual model training on the studied metrics, which is a natural next step; and (vii) the lack of direct quantification of the influence between C_m^* and text length, despite qualitative indications.

Ethics Statement

For this study, we utilized publicly available, open-source datasets and code in accordance with their respective usage licenses. Across all datasets, the essays were anonymized, with the exception of Diplomatrix-BR, where author data is public; however, even in this instance, we did not incorporate author-identifying information into any aspect of our research.

Generative AI usage

The authors used generative AI tools (Gemini 3 Pro) to assist with writing (orthographic correction, paraphrasing and language refinement) and code development. All AI-generated suggestions were verified by the authors.

Aknowledgements

This work was carried out at the Center for Artificial Intelligence (C4AI-USP), with support by the University of São Paulo, by the São Paulo Research Foundation (FAPESP) (grant #2019/07665-4) and by the IBM Corporation. The project was also supported by the Ministry of Science, Technology and Innovation, with resources of Law N. 8,248, of October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published as Residence in TIC 13, DOU 01245.010222/2022-44. Marcelo Finger was partly supported by the São Paulo Research Foundation (FAPESP) (grants 2023/00488-5 (SPIRA-BM), 2024/19144-7 (SPRINT)); and the National Council for Scientific and Technological Development (CNPq) (grant PQ 302963/2022-7). Felipe Ribas Serras was supported by IBM through a scholarship managed by FUSP (Support Foundation for the University of São Paulo) (Project 3541) and by a PPI-SOFTEX scholarship managed by FUSP (Project 3970). Denis Deratani Mauá was partly supported by FAPESP (grant 2022/02937-9) and CNPq (grant 305136/2022-4). Igor Silveira was supported by FAPESP (grant 2026/15768-1). This study was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior — Brasil (CAPES) — Finance Code 001.

References

Aissa, W., Amaro, R., Antunes, D., Bañeras-Roux, T., Baptista, J., Catala, A., Correia, L., François, T., Garcia, M., Izquierdo-Álvarez, M., Mamede, N., Martins, V., Neves, M., Ribeiro, E., Rey, S. R., and Vanzeveren, E. (2025). The iRead4Skills intelligent complexity analyzer. In Habernal, I., Schulam, P., and Tiedemann, J., editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 73–84, Suzhou, China. Association for Computational Linguistics.

Amorim, E. C. F. and Veloso, A. (2017). A multi-aspect analysis of automatic essay scoring for Brazilian Portuguese. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 94–102. Association for Computational Linguistics.

Barbosa, A., Silveira, I. C., and Mauá, D. D. (2025). An empirical analysis of large language models for automated cross-prompt essay trait scoring in brazilian portuguese. *Journal of the Brazilian Computer Society*, 31(1):857–870.

Bazelato, B. S. and Amorim, E. C. F. (2013). A bayesian classifier to automatic correction of portuguese essays. In *Conferência Internacional sobre Informática na Educação (TISE)*, volume 18, pages 779–782.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Biber, D. (1995). *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge University Press, 1st edition.

Biber, D. and Conrad, S. (2009). *Register, Genre, and Style*. Cambridge University Press, 1st edition.

Boquio, E. N. V. and Naval, Jr., P. C. (2024). Beyond canonical fine-tuning: Leveraging hybrid multi-layer pooled representations of BERT for automated essay scoring. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2285–2295, Torino, Italia. ELRA and ICCL.

Brideau, L. (2025). Correlation: Pearson, spearman, and kendall's tau. UVA Library StatLab, University of Virginia.

Cavalcanti, R., Casini, G., Assis, G., Real, L., Vianna, D., Mann, P., and Paes, A. (2025). Diplomatrix-br: Um corpus paralelo de redaçoes de autoria humana e de llms no concurso de diplomacia brasileira. In *Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, pages 192–205. SBC.

Chalegre, P. C., Machado, V. d. R., and Feltrim, V. D. (2026). Avaliação automática de redações do enem: Uma análise comparativa entre engenharia de características e transformers. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 738–748, Salvador, Brazil. Association for Computational Linguistics.

Croft, W. (2002). *Typology and Universals*. Cambridge University Press, 2 edition.

Ehret, K. (2021). An information-theoretic view on language complexity and register variation: Compressing naturalistic corpus data. *Corpus Linguistics and Linguistic Theory*, 17(2):383–410.

Ehret, K. and Szmrecsanyi, B. (2016). *An information-theoretic approach to assess linguistic complexity*, pages 71–94. De Gruyter, Berlin, Boston.

Fonseca, E. R., Medeiros, I., Kamikawachi, D., and Bokan, A. (2018). Automatically grading brazilian student essays. In *Computational Processing of the Portuguese Language. PROPOR 2018.*, pages 170–179.

Grünwald, P. D. and Vitányi, P. M. (2008). Algorithmic information theory. In Adriaans, P. and van Benthem, J., editors, *Philosophy of Information*, Handbook of the Philosophy of Science, pages 281–317. North-Holland, Amsterdam.

Hockett, C. (1958). *A Course in Modern Linguistics*. A Course in Modern Linguistics. Macmillan.

Hollander, M., Wolfe, D. A., and Chicken, E. (2014). *Nonparametric Statistical Methods*. John Wiley & Sons, Inc., Hoboken, New Jersey, 3rd edition.

Juola, P. (1998). Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics*, 5(3):206–213.

Juola, P. (2008). *Assessing linguistic complexity*, pages 89–108. John Benjamins Publishing Company.

Kendall, M. G. (1945). The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251.

Laerd Statistics (2018). Kendall's tau-b using spss statistics - a how-to statistical guide.

Leal, S. E., Duran, M. S., Scarton, C. E., Hartmann, N. S., and Aluísio, S. M. (2024). NILC-Metrix: Assessing the complexity of written and spoken language in Brazilian Portuguese. *Language Resources and Evaluation*, 58(1):73–110.

Leal, S. E., Serras, F. R., Finger, M., and Aluísio, S. M. (2026). Complexidade textual e suas tarefas relacionadas. In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, volume 3, book chapter 6. BPLN, 4 edition.

Liu, H. (2015). Comparing Welch's ANOVA, a Kruskal-Wallis test and traditional ANOVA in case of heterogeneity of variance. Master's thesis, Virginia Commonwealth University, Richmond, VA. VCU Theses and Dissertations.

Liu, J., Xu, Y., and Zhu, Y. (2019). Automated essay scoring based on two-stage learning.

Liu, Y., Zhang, Z., Zhang, W., Yue, S., Zhao, X., Cheng, X., Zhang, Y., and Hu, H. (2023). Argugpt: evaluating, understanding and identifying argumentative essays generated by gpt models.

Lobo, T. R. and Martins, C. A. (2026). Evolução de padrões linguísticos na escrita científica em português: Uma análise com NILC-metrix. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of*

the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1, pages 1062–1067, Salvador, Brazil. Association for Computational Linguistics.

Locatelli, M. S., Miranda, M. P., Da Silva Costa, I. J., Prates, M. T., Thomé, V., Zaparoli Monteiro, M., Lacerda, T., Pagano, A., Rios Neto, E., Meira Jr., W., and Almeida, V. (2024). Examining the behavior of llm architectures within the framework of standardized national exams in brazil. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7:879–890.

Marinho, J. C., Anchiêta, R. T., and Moura, R. S. (2021). Essay-br: a brazilian corpus of essays. In *Anais do III Dataset Showcase Workshop*, pages 53–64. Sociedade Brasileira de Computação.

Marinho, J. C., Cordeiro, F., Anchiêta, R. T., and Moura, R. S. (2022). Automated essay scoring: An approach based on enem competencies. In *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional*, pages 49–60. SBC.

Matos, G. G. d. and Feltrim, V. D. (2026). Avaliação automática de redações do enem: Um estudo empírico sobre representações linguísticas e contextuais. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 488–497, Salvador, Brazil. Association for Computational Linguistics.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., and Cai, Z. (2014). *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge University Press, 1st edition.

McWhorter, J. H. (2001). The worlds simplest grammars are creole grammars. *Linguistic Typology*, 5(2-3).

Mello, R. F., Oliveira, H., Wenceslau, M., Batista, H., Cordeiro, T., Bittencourt, I. I., and Isotanif, S. (2024). Propor'24 competition on automatic essay scoring of portuguese narrative essays. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 2*, pages 1–5.

Monteiro, R., Correia, S., Amaro, R., Moutinho, M., Barbosa, S., and Reis, M. L. (2025). Níveis e descritores de complexidade textual para adultos de baixa literacia: um referencial do projeto iread4skills. *Revista da Associação Portuguesa de Linguística*, (13):193–222.

Nichols, J. (1998). *Linguistic Diversity in Space and Time*. University of Chicago Press.

Page, E. B. (1966). The imminence of... grading essays by computer. *The Phi Delta Kappan*, pages 238–243.

Rossman, L. N., Silveira, I. C., and Mauá, D. D. (2026). Evaluating automated scoring models on official ENEM essays. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 161–171, Salvador, Brazil. Association for Computational Linguistics.

Sardinha, T. B., Kauffmann, C., and Acunzo, C. M. (2014). A multi-dimensional analysis of register variation in Brazilian Portuguese. *Corpora*, 9(2):239–271.

Serras, F., Carpi, M., Branco, M., and Finger, M. (2024). Analysing and validating language complexity metrics across South American indigenous languages. In Kuribayashi, T., Rambelli, G., Takmaz, E., Wicke, P., and Oseki, Y., editors, *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 152–165, Bangkok, Thailand. Association for Computational Linguistics.

Serras, F. R., Carpi, M. D. M., Sturzeneker, M. L., Palma, M. F., Costa, A. S., Monte, V. M. D., Namiuti, C., Crespo, M. C. R. M., Paixão De Sousa, M. C., and Finger, M. (2026). Análise e Classificação Automática de Domínios Discursivos no Português do Brasil. *Linguamática*, 17(2):131–171.

Serras, F. R. and Finger, M. (2026). Compression-based language complexity under register variation in Portuguese. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 808–818, Salvador, Brazil. Association for Computational Linguistics.

Silveira, I. C., Barbosa, A., da Costa, D. S. L., and Mauá, D. D. (2025a). Investigating universal adversarial attacks against transformers-based automatic essay scoring systems. In Paes, A. and Verri, F. A. N., editors, *Intelligent Systems*, pages 169–183, Cham. Springer Nature Switzerland.

Silveira, I. C., Barbosa, A., and Mauá, D. D. (2024). A new benchmark for automatic essay scoring in Portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 228–237.

Silveira, I. C. and Mauá, D. D. (2026). Neuro-symbolic approaches for rubric-based automatic essay evaluation of ENEM essays. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 790–799, Salvador, Brazil. Association for Computational Linguistics.

Silveira, I. C., Ribeiro, E., Mamede, N., and Baptista, J. (2025b). Aprendizado por transferência para correçao automática de redaçao. *Linguamática*, 17(2):99–116.

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, 38(3/4):330–336.

Validating Semantic Content in Sentence Simplification using OpenIE and Relational Graphs

Samuel Rios da Silva[1], **Larrissa Dantas**[1], **Daniela Barreiro Claro**[1], **Aline Paes**[2], **Maria José Finatto**[3], **Rerisson Cavalcante**[4], **Silvana Ribeiro**[4]

[1]FORMAS - Centro de Pesquisa em Dados e Linguagem Natural
Instituto de Computação, Universidade Federal da Bahia (UFBA)
Salvador, BA, Brasil
{samuelrs, larrissasilva, dclaro}@ufba.br

[2]Instituto de Computação, Universidade Federal Fluminense (UFF)
Niterói, RJ, Brasil
alinepaes@ic.uff.br

[3]Universidade Federal do Rio Grande do Sul
Porto Alegre, RS, Brasil
mariafinatto@gmail.com

[4]Instituto de Letras, Universidade Federal da Bahia
Salvador, BA, Brasil
{rerisson.cavalcante, silvanar}@ufba.br

Abstract. *Sentence simplification often alters sentence structure while aiming to preserve meaning, making semantic fidelity difficult to evaluate. In this paper, we investigate whether graph-based representations derived from Open Information Extraction (Open IE) can capture semantic preservation across different simplification levels. Using the PorSimplesSent corpus, we extract relational triples from original and simplified sentences to build semantic graphs for comparison. Our results provide exploratory evidence that graph-based analysis can capture meaningful semantic differences and provides complementary signals beyond surface-level lexical overlap for evaluating simplification quality.*

Resumo. *A simplificação de sentenças frequentemente altera a estrutura textual enquanto preserva o significado, tornando a fidelidade semântica difícil de avaliar. Neste trabalho, investigamos se representações baseadas em grafos derivadas de Extração Aberta de Informação (Open IE) podem capturar a preservação semântica em diferentes níveis de simplificação. Utilizando o corpus PorSimplesSent, extraímos triplas relacionais de sentenças originais e simplificadas para construir grafos semânticos comparáveis. Os resultados fornecem indícios exploratórios de que a análise baseada em grafos pode capturar diferenças semânticas relevantes e fornece sinais complementares além da sobreposição lexical superficial para avaliar a qualidade da simplificação.*

1. Introdução

A simplificação textual tem ganhado relevância crescente no contexto brasileiro, especialmente após a promulgação da Lei nº 15.263/2025, que instituiu o uso de linguagem simples em documentos e comunicações públicas[1]. Essa iniciativa buscou ampliar a acessibilidade às informações prestadas por órgãos públicos e garantir que seus conteúdos

[1] https://www2.camara.leg.br/legin/fed/lei/2025/lei-15263-14-novembro-2025-798293-publicacaooriginal-177011-pl.html

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 405–419
October 19-22, 2026 ©2026 Association for Computational Linguistics

fossem compreendidos por uma parcela mais ampla da população, incluindo pessoas com diferentes níveis de letramento e proficiência linguística, e pessoas com necessidades especiais [Finatto 2024].

Nesse cenário, a área de Processamento de Linguagem Natural (PLN) tem desempenhado um papel central no desenvolvimento de métodos automáticos voltados à simplificação textual [Leal et al. 2024, Leal et al. 2026]. Paralelamente, diversos estudos têm se dedicado à construção de recursos e insumos voltados à avaliação automática da complexidade textual e dos níveis de legibilidade [de Menezes et al. 2023].

As técnicas contemporâneas de simplificação automática de textos combinam abordagens linguísticas, métodos de aprendizado de máquina e, mais recentemente, modelos de linguagem de grande porte (LLMs) [Kim 2022, Liu et al. 2024, Alves et al. 2023, Scalercio et al. 2024, Shardlow et al. 2024, Baez and Saggion 2023, Scalercio et al. 2025]. Tradicionalmente, essas abordagens utilizam conjuntos de características linguísticas e corpora paralelos, compostos por textos complexos e suas respectivas versões simplificadas, para orientar modelos supervisionados ou estratégias heurísticas de simplificação [Chandrasekar et al. 1996b]. As transformações podem ocorrer em diferentes níveis linguísticos, desde o nível lexical, com a substituição de termos complexos por alternativas mais frequentes ou mais acessíveis, até o nível sintático, envolvendo a divisão de sentenças, a redução de constituintes e a reorganização de estruturas para formas mais diretas e canônicas [Saggion and Hirst 2017].

Um *corpus* paralelo pioneiro que oferece acervos de simplificações de frases para o português do Brasil é o PorSimplesSent [Leal et al. 2018]. O PorSimplesSent utiliza técnicas de simplificação de sentenças baseadas em aspectos lexicais e sintáticos e detém uma estrutura composta por uma sentença original complexa (Original), acompanhadas de duas versões simplificadas. A primeira versão sofreu apenas simplificação lexical e é denominada Natural. A segunda versão (chamada *Strong*) emprega técnicas sintático-lexicais e inclui elementos psicolinguísticos, com a incorporação e/ou o resgate de elementos mencionados em sentenças anteriores.

Nesse processo, preservar o significado da sentença original em suas versões simplificadas permanece um dos principais desafios da simplificação automática de textos [Agrawal and Carpuat 2024]. Embora as transformações realizadas possam melhorar aspectos como legibilidade e acessibilidade, elas também podem introduzir perdas de informação, omissões relevantes ou alterações semânticas sutis, comprometendo a fidelidade do conteúdo original. Nesse contexto, diferentes trabalhos, sobretudo na língua inglesa, têm investigado métodos para avaliar a preservação semântica na simplificação textual [Vásquez-Rodríguez et al. 2021].

A título ilustrativo, segue um exemplo de simplificação com omissão de informação (omissão que a ópera *estava moribunda*), acréscimo e/ou retomada de elementos (para o item *dele*, referido como *Pavarotti* em sentenças anteriores) e alteração da ordem de elementos (substantivo-adjetivo - ópera viva)[2].

- **Original**: A importância dele é ter mantido viva a ópera, que estava moribunda – diz Chaves.

[2]Sentença do extrato aleatório do *corpus* PorSimplesSent.

- **Simplificada (neutra)**: Chaves diz que a importância de Pavarotti é ter mantido viva a ópera.
- **Simplificada (strong)**: Chaves diz que a importância de Pavarotti é ter mantido a ópera viva.

A avaliação automática em simplificação textual ainda representa um desafio, especialmente no que se refere à preservação semântica, uma vez que métricas tradicionais frequentemente não capturam adequadamente omissões de informação, alterações de significado ou reorganizações estruturais mais profundas [Alva-Manchego et al. 2021]. Por outro lado, embora avaliações humanas sejam capazes de identificar tais fenômenos com maior precisão, sua realização em larga escala torna-se impraticável devido ao alto custo de anotação, à necessidade de especialistas e ao tempo requerido para análises consistentes. Dessa forma, torna-se necessária a investigação de novas formas de avaliar a simplificação textual, principalmente no que cerne a preservação da semântica.

Com isso em mente, o presente trabalho realiza uma análise qualitativa das sentenças originais em relação às respectivas simplificações que integram o *corpus* PorSimplesSent, por meio de grafos gerados a partir de informações extraídas das sentenças. Partimos da hipótese de que o significado central de um texto pode ser representado por relações semânticas explícitas entre entidades e predicados. Assim, ao converter sentenças originais e simplificadas em grafos relacionais, torna-se possível comparar não apenas a similaridade lexical superficial, mas também a preservação da estrutura semântica subjacente. Divergências estruturais entre os grafos podem indicar perda de informação, omissão de relações relevantes ou alterações semânticas introduzidas pela simplificação, enquanto similaridades sugerem maior fidelidade ao conteúdo original.

Para investigar essa hipótese de forma exploratória, triplas de sentenças foram extraídas do *corpus* PorSimplesSent. O principal objetivo foi explorar como a estrutura relacional dos grafos gerados reflete a preservação semântica em diferentes níveis de simplificação. Com auxílio desses grafos e da apreciação de avaliadores humanos, descrevemos em que medida esse método pode apoiar a verificação da preservação semântica em diferentes processos de simplificação.

Especificamente, pretende-se responder às seguintes questões de pesquisa:

- Como os diferentes níveis e estratégias de simplificação presentes no *corpus* PorSimplesSent se refletem na estrutura dos grafos gerados por Open IE?
- Em casos de preservação semântica, de que forma essa conservação se manifesta na estrutura relacional dos grafos?
- Em casos de distorção, omissão ou alteração semântica, como essas modificações impactam a estrutura dos grafos gerados?

Os experimentos foram divididos em quatro etapas principais. A primeira etapa analisou a omissão de informação e a mudança de sentido com a participação de avaliadores linguistas. A segunda buscou observar se os grafos gerados a partir das sentenças simplificadas detinham graus menores de complexidade que os grafos de sentenças originais. A terceira etapa buscou descrever em que medida tanto omissões de informação quanto mudanças de sentido repercutem na geração de novos nós e arestas nos grafos das sentenças simplificadas. Na quarta etapa, foram comparadas as anotações realizadas pelos linguistas com os resultados obtidos a partir dos grafos gerados.

O presente artigo está organizado como segue. A Seção 2 apresenta os trabalhos relacionados. A seção 3 descre a tarefa de simplificação. A Seção 4 descreve a proposta deste trabalho. A Seção 5 descreve os experimentos e resultados, e casos particulares são discutidos na Seção 6. Por fim, a Seção 7 apresenta as conclusões e os trabalhos futuros.

2. Trabalhos Relacionados

Diversos trabalhos investigam os desafios do processo de avaliação da simplificação textual, seja por meio de métricas automáticas, de análise morfossintática ou ainda de avaliação humana.

Métricas Automáticas e Morfossintáticas: As métricas automáticas baseadas em referências, como SARI [Xu et al. 2016] e BERTScore [Zhang et al. 2020], são utilizadas na literatura para medir simplicidade e preservação de significado [Scalercio et al. 2025]. Para avaliações sem referência, índices clássicos de legibilidade como o Flesch Readability Index e o Flesch-Kincaid Grade Level (FKGL) [Kincaid et al. 1975] ainda são adotados para verificar a legibilidade da sentença resultante [Scalercio et al. 2026, Alva-Manchego et al. 2020]. Métricas morfossintáticas baseadas em hipóteses psicolinguísticas [Heineman et al. 2023] têm sido utilizadas para validar o processo de simplificação, como a razão Lemma/Token para diversidade lexical, a proporção de voz passiva para ativa, proporção de orações adverbiais que precedem a oração principal, capturando tendências na estrutura da frase, e razão entre orações relativas totalmente desenvolvidas e reduzidas, refletindo simplificações sintáticas [Scalercio et al. 2025, Scalercio et al. 2026].

Avaliação Humana e Análise de Erros: Diante da insuficiência das métricas automáticas na simplificação de sentenças, a avaliação humana permanece como o padrão-ouro. Abordagens tradicionais analisam o texto simplificado sob uma escala Likert de três eixos fundamentais: adequação (preservação do significado), simplicidade (as sentenças geradas são mais simples?) e fluência (as frases geradas são gramaticais ou bem formadas?) [Alva-Manchego et al. 2020]. [Wu and Arase 2026] propõe uma abordagem focada em falhas e voltada ao resultado, mapeando sete tipos de erros (como Altered Meaning e Lack of Simplicity, subdivididos em eixos lexicais e estruturais).

Embora alguns métodos propostos [Ivanov and Solnyshkina 2020, Silva et al. 2025, Queiroz et al. 2023] utilizem grafos derivados de representações semânticas, nenhum dos trabalhos mapeados investiga a preservação do conteúdo semântico por meio da extração de informação aberta e da análise de divergências estruturais entre grafos gerados a partir de diferentes níveis de simplificação.

3. Simplificação de Sentenças

A tarefa de simplificação de sentenças consiste em transformar uma sentença de entrada, tipicamente complexa, em uma ou mais sentenças de saída linguisticamente mais simples, com o objetivo de facilitar sua leitura e compreensão, preservando ao máximo seu conteúdo semântico original [Chandrasekar et al. 1996a, Al-Thanyyan and Azmi 2021].

A simplificação pode envolver diferentes tipos de operações, incluindo substituições lexicais por termos mais frequentes ou acessíveis, reestruturação sintática por meio da decomposição de sentenças longas e complexas em unidades mais curtas, e reorganização discursiva para melhorar a fluidez e a clareza textual.

Tipicamente, um *corpus* de simplificação, como o PorSimplesSent [Leal et al. 2018], inclui pares alinhados entre uma sentença original e suas versões simplificadas em diferentes níveis de intensidade, permitindo investigar sistematicamente os efeitos de simplificação na preservação semântica.

Dentre o processo de simplificação das sentenças no PorSimplesSent, há sentenças que são simplificadas por uma operação de *split*. Consequentemente, a simplificação de uma sentença original pode resultar em duas ou mais sentenças simplificadas. Como a simplificação por *split* gera triplas e grafos independentes para cada nova sentença, a comparação estrutural entre versões simplificadas e originais torna-se mais difícil e menos interpretável. Por essa razão, foram selecionadas apenas sentenças cuja simplificação não envolvesse processos de *split*.

4. Análise semântica da simplificação

Com o intuito de validar o conteúdo semântico na simplificação de sentenças por meio de extração de informação [Oliveira et al. 2022] e seus respectivos grafos, a metodologia foi organizada em quatro etapas principais: (1) a preparação do corpus e a representação linguística das sentenças; (2) a análise semântica das sentenças simplificadas sob a perspectiva da omissão de informação e preservação da semântica; (3) análise da simplificação via extração de informação e seus grafos, e, por fim, (4) uma análise comparativa quanto aos grafos gerados, considerando limitações em sentenças complexas e simplificações com mudança semântica significativa. A Figura 1 descreve o pipeline de validação da preservação do conteúdo semântico das sentenças simplificadas.

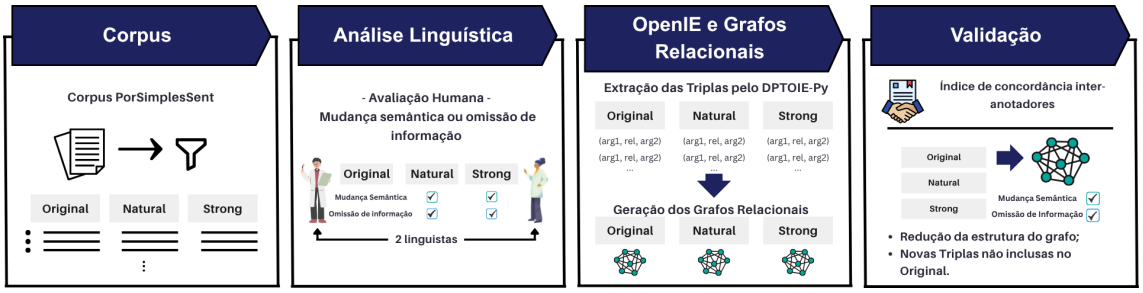


Figura 1. Pipeline

A **primeira etapa** consistiu na preparação do corpus a ser utilizado na validação da simplificação. Considerando os dois níveis de simplificação: $O \rightarrow N$ (**O**riginal para **N**atural) e $O \rightarrow S$ (**O**riginal para **S**trong), foi aplicado um filtro no *corpus* PorSimplesSent [Leal et al. 2018] para selecionar os pares de sentenças que não sofreram o processo de *split* e foram modificadas no processo de simplificação (i.e. cuja sentença original se difere léxico-sintaticamente da sentença simplificada), resultando em um total de 264 sentenças.

Desse conjunto, foram selecionadas 20 sentenças originais alinhadas com suas versões simplificadas, objetivando obter um conjunto diverso de representações linguísticas do PorSimplesSent para análise e validação por meio deste trabalho. O processo de aleatoriedade foi feito selecionando 20 valores no intervalo de [1, 264].

A **segunda etapa** foi responsável pela análise semântica das sentenças simplificadas sob a perspectiva da omissão da informação e da mudança de sentido. As 20 sentenças

selecionadas foram avaliadas manualmente por dois linguistas, responsáveis por analisar, independentemente, cada simplificação, observando dois critérios definidos:

- *Omissão de informação:* a omissão de informação resulta na remoção de conteúdos relevantes das sentenças simplificadas que estavam presentes na sentença original.
- *Mudança de sentido:* a mudança de sentido corresponde às alterações semânticas que alteram o significado da sentença original.

A **terceira etapa** consistiu na definição do método de extração de triplas a ser utilizado e na geração dos grafos. Cada sentença s, em sua versão Original ou Simplificada, foi processada através do método de extração de informação denominado *DPToie-Py*[3] e gerou um conjunto de triplas $T(s) = \{(arg_1, rel, arg_2), ...\}$, onde $T(s)$ representa a função que mapeia uma sentença para um conjunto de triplas relacionais [Oliveira et al. 2022].

Associamos à sentença s o grafo $G(s) = (V_s, E_s)$, onde o conjunto de vértices V_s é gerado através dos argumentos das triplas ($arg_1, arg_2 \in V_s$) e o conjunto de arestas E_s é gerado através da relação rel de seus argumentos. $G(s)$, portanto, é a função de construção do grafo da sentença s a partir de triplas extraídas em $T(s)$ [Oliveira et al. 2022].

Com o intuito de dar um maior subsídio à análise dos grafos, decorrente do processo de simplificação textual, a árvore de dependência de cada sentença $UD(s) = PortParser(s)$ também foi gerada através do Portparser.v2 [Lopes and Pardo 2024]. Assim, a análise do processo de simplificação é representada através de cada sentença s e seus elementos associados $\beta_s = \{T(s), G(s), UD(s)\}$ nas versões **O**riginal, **N**atural e **S**trong.

A hipótese central é a de que simplificações de sentenças que preservam a semântica tendem a manter a mesma estrutura informacional do grafo original, ainda que apresentem redução estrutural. A validação dos grafos considerou dois aspectos principais:

- *Redução estrutural do grafo:* análise da simplificação topológica decorrente da redução textual, observando diminuição no número de nós, arestas ou relações;
- *Introdução de novas informações:* identificação de triplas inexistentes na sentença original, indicando possíveis distorções semânticas, inferências indevidas ou mudanças de sentido decorrentes da simplificação.

Por fim, na **quarta etapa**, foram comparadas as anotações dos linguistas aos resultados obtidos a partir dos grafos gerados. Essa comparação teve como objetivo verificar o grau de correspondência entre a avaliação humana de preservação semântica na simplificação de sentenças e os indícios estruturais observados nos grafos de conhecimento gerados a partir destas sentenças.

5. Experimentos e Resultados

Esta seção descreve os experimentos conduzidos neste trabalho e discute os resultados obtidos.

[3]Versão melhorada do DPToie[Oliveira et al. 2022] em Python, com melhorias na extração de relações semânticas

5.1. Experimento A - Análise Linguística

Considerando a efetiva redução de complexidade observada nas sentenças simplificadas de acordo com o PorSimplesSent, foram adotados dois critérios de avaliação com o objetivo de analisar o comportamento das simplificações em termos de *omissão de informação* e *preservação da semântica*. Esses critérios são descritos a seguir:

- *Omissão de informação:* a omissão de informação é responsável pela remoção de conteúdo relevante nas sentenças simplificadas, que estavam presentes na sentença original.
- *Mudança de sentido:* a mudança de sentido identifica as alterações semânticas que modifiquem o significado da sentença original.

Cada conjunto de sentenças **O**riginal, **N**atural e **S**trong foi analisado sob esses dois critérios por dois linguistas seniores, que realizaram as suas avaliações de maneira independente.

5.1.1. Resultados

Os resultados foram divididos em duas etapas de avaliação da simplificação: $O \rightarrow N$ e $O \rightarrow S$. Em relação à *omissão de informação*, das 20 sentenças simplificadas, os anotadores apresentaram concordância de 80% para a classe negativa, ou seja, não há *omissão de informação* nas simplificações, tanto para as $O \rightarrow N$ quanto $O \rightarrow S$. Somente em um caso, ambos concordaram com a *omissão da informação* nas sentenças, como segue na Tabela 1.

Versão	**Sentença**
Original	A importância dele é ter mantido viva a ópera, que estava moribunda – diz Chaves.
Natural	Chaves diz que a importância de Pavarotti é ter mantido viva a ópera.

Tabela 1. Exemplo de simplificação textual com *omissão de informação* entre as versões $O \rightarrow N$ com concordância inter-anotadores.

Em relação à *mudança de sentido*, observa-se que, na simplificação $O \rightarrow N$, a concordância observada inter-anotadores é 70%, sendo quatro casos de concordância que houve *mudança de sentido* na frase simplificada em relação à original e 10 casos de concordância para a classe negativa. Além disso, há seis casos nos quais o Anotador 2 considerou que houve *mudança de sentido* em divergência do Anotador 1 (Tabela 2).

Versão	**Sentença**
Original	Segundo eles, a pele humana ainda é bem mais complexa do que as imitações desenvolvidas até agora.
Natural	Eles dizem que a pele humana ainda é mais muito mais complexa do que as cópias desenvolvidas até agora.

Tabela 2. Exemplo de simplificação textual com mudança de sentido entre as versões $O \rightarrow N$ com concordância inter-anotadores.

Ainda em relação à *mudança de sentido*, observa-se que em relação às sentenças simplificadas das versões Strong, a concordância observada inter-anotadores decresceu

para 65%, sendo cinco casos de concordância que houve mudança de sentido na frase simplificada em relação à original. Além disso, há 7 casos nos quais o Anotador 2 considerou que houve *mudança de sentido*, mas esse comportamento não foi observado pelo Anotador 1.

Observa-se que, na avaliação humana, não houve divergência quanto à *omissão de informação* nas sentenças simplificadas. Por outro lado, há divergência entre anotadores quanto à preservação da semântica considerando a análise da *mudança de sentido*, definindo assim a tarefa como difícil. Nestes casos, evidencia-se a necessidade de analisar se esses comportamentos se refletem nos grafos gerados por meio das triplas.

5.2. Experimento B - Análise dos Grafos

Esse experimento analisa a simplificação sob a perspectiva dos grafos relacionais. Para cada sentença s dos pares $O \rightarrow N$ e $O \rightarrow S$ foi extraído um conjunto $T(s)$ de triplas (ver Tabelas 3 e 4) e, consequentemente, gerado o grafo $G(s)$ por meio das triplas extraídas (ver Figuras 2 e 3).

A avaliação da simplificação das sentenças por meio dos grafos teve como principal objetivo analisar as relações entre as omissões de informações e a estrutura do grafo, assim como as relações entre as mudanças de sentido nas sentenças simplificadas e as modificações nas arestas e nós dos grafos das sentenças simplificadas.

A análise de redução de um grafo foi realizada independentemente da omissão de informação, com o intuito de verificar se, mesmo na ausência de omissão, ocorre redução do grafo na simplificação de sentenças. Consequentemente, a observação de novos nós e arestas também foi analisada independentemente com o intuito de isolar os casos de preservação de semântica, mesmo em caso de não omissão de informação. As árvores de dependência $UD(s_i)$ foram utilizadas para ampliar a compreensão das extrações e grafos gerados.

5.2.1. Resultados

Considerando a análise de redução do grafo nas sentenças simplificadas $O \rightarrow N$, observou-se que apenas quatro grafos de sentenças simplificadas reduziram sua estrutura topológica, o que corresponde a uma redução de 20%. Em relação aos grafos simplificados $O \rightarrow S$, esse número aumentou para sete, perfazendo um total de 35% das simplificações. Neste caso, embora haja uma simplificação das sentenças, elas não são refletidas na estrutura topológica dos grafos.

Avaliando a correspondência com a *omissão de informação*, no único caso em que ambos os anotadores concordaram (Tabela 3), também não houve redução na estrutura do grafo (Figura 2).

Considerando a *preservação da semântica*, ou seja, que no processo de simplificação não houve *mudança de sentido* das sentenças simplificadas $O \rightarrow N$ e $O \rightarrow S$, os grafos gerados para cada uma delas foram avaliados quanto à existência de uma alteração nos nós e nas arestas, caracterizando novos fatos incorporados que não existiam na versão original da sentença.

Observou-se que, em 30% das simplificações $O \rightarrow N$, houve alterações nas arestas e nós dos grafos simplificados. Considerando as quatro concordâncias entre anotadores sobre a *mudança de sentido* de $O \rightarrow N$, em 50% também houve uma correspondência na modificação das arestas e dos nós. Os dois casos específicos nos quais os anotadores concordaram que houve uma mudança de sentido, mas não houve novo nó ou aresta, observou-se que o tema foi *política*, e a concordância de ambos se referiu ao significado do termo "união" que diverge de "coligação". Esse resultado não tem impacto direto no grafo gerado, conforme a Figura 3.

	Original			**Natural**		
#	arg_1	rel	arg_2	arg_1	rel	arg_2
1	A importância de ele	é	mantido	A importância de Pavarotti	é	mantido
2	A importância de ele	é	mantido viva a ópera , estava moribunda	A importância de Pavarotti	é	mantido viva a ópera
3	A importância de ele	ter mantido	viva a ópera , estava moribunda	A importância de Pavarotti	é	viva a ópera
4				A importância de Pavarotti	ter mantido viva	a ópera
5				Chaves	ter mantido	viva a ópera, estava moribunda

Tabela 3. Conjuntos de triplas da sentença Original 12 e de sua versão Simplificada ($O \rightarrow N$).

Considerando as simplificações $O \rightarrow S$, em 40% dos casos houve *mudança de nós e arestas* nos grafos gerados. Nestes casos, dos cinco casos de concordância entre anotadores com *mudança de sentido* $O \rightarrow S$, em 80% dos casos houve alteração nas arestas e nós. Neste caso, há um forte indício de que o grafo reflete a mudança de sentido, quando há uma concordância entre os anotadores.

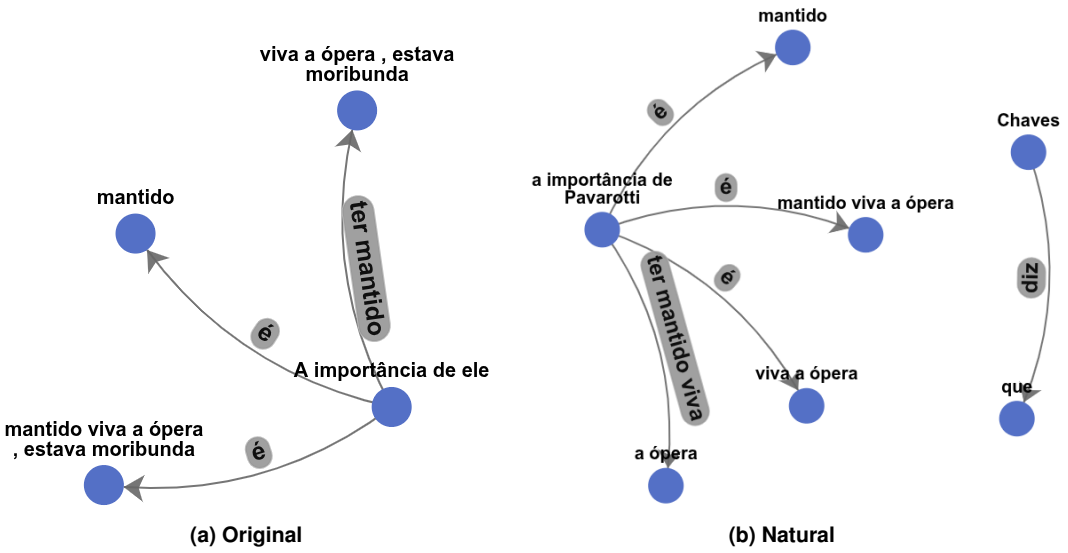


Figura 2. Representação em grafo da simplificação da sentença Original na versão Natural (Sentença 12).

Observou-se que a omissão de informação não foi uma das causas da mudança de sentido neste *corpus*, visto que poucas foram as ocorrências de omissões (1 de $O \rightarrow N$ e 3 de $O \rightarrow S$).

Em sua maioria, a mudança de sentido ocorreu em um único nó, com a substituição do termo por outro termo que não carregava mais o mesmo significado semântico, tais como "união" e "coligação", "devota" e "apegada". Nestes casos, não há mudança no grafo e nem essas nuances no significado são absorvidas para gerar novas arestas e novos nós.

	Original			**Natural**		
#	arg_1	rel	arg_2	arg_1	rel	arg_2
1	A regra	obriga fechar	apenas coligações não colidam com as nacionais	A regra	obriga fechar	apenas as uniões não se choquem com as nacionais
2	A regra	obriga fechar	em os Estados apenas coligações não colidam com as nacionais	A regra	obriga fechar	em os Estados apenas coligações as uniões não se choquem com as nacionais
3	A regra	obriga fechar	em os Estados	A regra	obriga fechar	em os Estados

Tabela 4. Conjuntos de triplas da sentença Original 14 e de sua versão Simplificada ($O \rightarrow N$).

6. Discussões

As discussões refletem algumas particularidades durante o processo de avaliação da geração de grafos a partir da Open IE para a validação da preservação semântica na simplificação de sentenças.

Nos casos em que houve a redução estrutural, não houve modificações nos grafos, ou seja, arestas novas ou nós novos não foram criados.

Os resultados indicam que a análise de preservação semântica por meio de grafos gerados por OpenIE fornece bons indícios de alterações semânticas, principalmente quando há uma concordância entre os anotadores. Essa concordância reflete que a mudança de sentido é observável. Além disso, quando a mudança ocorre no nível da sintaxe, envolvendo principalmente o deslocamento do "root" na Árvore de Dependência, os novos nós ou arestas nos grafos gerados por esse método de OpenIE tende a ser mais evidente, devido à heurística intrínseca ao método.

As mudanças no significado de um ou dois termos em uma sentença, tais como as diferenças políticas entre "união" e "coligação", ou ainda casos como "devoto" e "apegado" não são refletidas em novas arestas ou nós no grafo, visto que estão inseridas no arg_1 ou arg_2 (Figura 3). Nestes casos, o grafo não reflete essa mudança de significado. Por outro lado, em se tratando de aposto, ou seja, uma omissão de informação da sentença original, o grafo reflete essa omissão tanto topologicamente, quando na mudança de sentido, caso o aposto seja permutado como agente da sentença.

7. Conclusões

Este trabalho investigou o uso de representações baseadas em grafos derivados de técnicas de Open IE, para apoiar a análise da preservação semântica em tarefas de simplificação

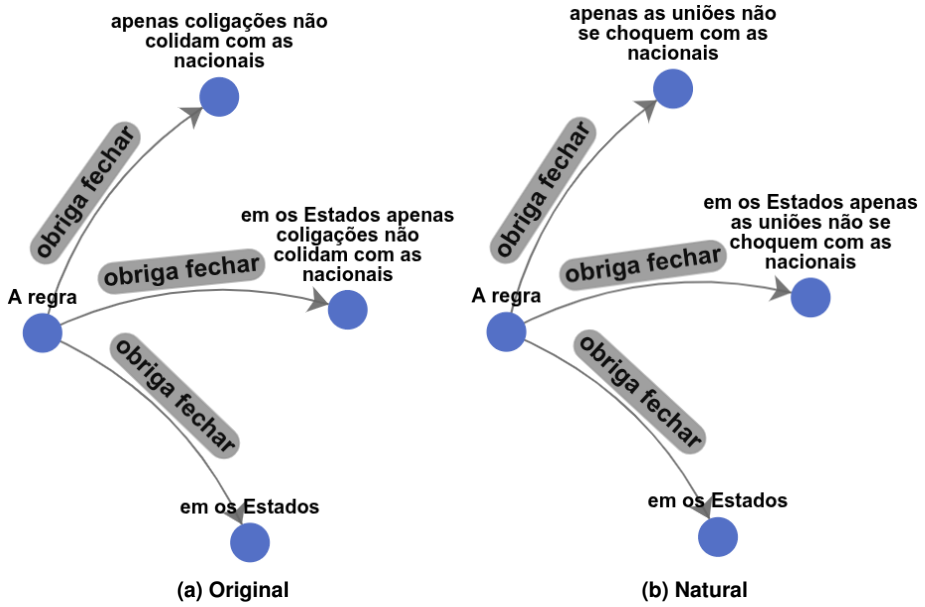


Figura 3. Representação em grafo da simplificação da sentença $O \rightarrow S$ (Sentença 14) com mudança de sentido mas sem mudança no grafo.

de sentenças em português. A partir do corpus PorSimplesSent, foram comparadas as versões original e simplificadas, com o objetivo de verificar em que medida as alterações de complexidade textual se refletem em mudanças estruturais nos grafos gerados.

Os resultados indicam que os grafos derivados de Open IE fornecem indícios úteis sobre a preservação ou a alteração de significado na simplificação textual. Em particular, divergências estruturais entre os grafos da sentença original em relação às simplificadas mostraram-se correlacionadas aos casos de mudança de sentido, conforme identificado na análise linguística. As omissões de informação não foram observadas sendo refletidas nas estruturas topológicas dos grafos.

Como trabalho futuro, pretendemos ampliar a avaliação para outros corpora, diferentes métodos de extração de informação e novos domínios textuais, investigando se os achados deste estudo exploratório permanecem consistentes em cenários mais amplos.

Limitações

Nesta seção descrevemos as limitações evidenciadas em relação a este trabalho. Algumas extrações, de acordo com as heurísticas definidas pelo DPToie-Py, não foram realizadas, e, consequentemente, os grafos não foram gerados, resultando em grafos vazios (Sentenças *Strong 17* e *Strong 9*). Essa limitação impediu a análise nestes casos. Outra limitação deste trabalho é em relação ao subconjunto do *corpus* PorSimplesSent, definido como aleatório, não contemplando uma amplitude de casos linguísticos por tipos de simplificação, condicionando os experimentos a uma baixa cobertura para análise dos grafos. A expansão do *corpus* e a definição de um guia de anotação são trabalhos futuros para mitigar essas limitações.

Declaração Ética

Este trabalho investiga a avaliação da preservação semântica na simplificação textual, utilizando representações relacionais derivadas de *Open Information Extraction* (Open IE). Como a pesquisa foi conduzida sobre um corpus público de simplificação textual, não houve coleta de dados pessoais sensíveis nem participação direta de seres humanos no processo experimental. Entretanto, reconhece-se que tanto os sistemas de simplificação automática quanto os métodos de extração de informação podem introduzir vieses linguísticos, omissões semânticas ou interpretações incorretas das relações presentes no texto original. Tais limitações podem impactar especialmente grupos com diferentes níveis de letramento, bem como aplicações em domínios sensíveis, como educação, saúde e comunicação governamental. Além disso, as representações relacionais produzidas por Open IE dependem da qualidade das extrações automáticas, podendo refletir erros do sistema de extração e não apenas alterações semânticas introduzidas pela simplificação. Assim, os resultados deste trabalho devem ser interpretados como sinais auxiliares de preservação semântica, e não como uma medida definitiva de equivalência de significado.

Uso de IA generativa

Ferramentas de IA generativa foram utilizadas para o refinamento textual e a correção gramatical.

Agradecimentos

Os autores agradecem o financiamento parcial da bolsa CAPES do PIBIC 2025/2026, do apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001 (processo nº 88887.970806/2024-00), do CNPq (processo 307088/2023-5) e da FAPERJ (processos SEI-260003/002930/2024 e SEI-260003/000614/2023). Este trabalho também conta com o apoio do Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação (INCT-TILD-IAR) (processo 408490/2024-1). E por fim, os autores agradecem o apoio do CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), por meio da bolsa de Produtividade em Pesquisa (PQ, processo 307088/2023-5), do PROBIC FAPERGS-UFRGS e da FAPERGS, por meio do Edital 06/2025 – Programa Pesquisador Gaúcho (PqG) (processo 25/2551-0002702).

Referências

Agrawal, S. and Carpuat, M. (2024). Do text simplification systems preserve meaning? a human evaluation via reading comprehension. *Transactions of the Association for Computational Linguistics*, 12:432–448.

Al-Thanyyan, S. S. and Azmi, A. M. (2021). Automated text simplification: a survey. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

Alva-Manchego, F., Scarton, C., and Specia, L. (2020). Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.

Alva-Manchego, F., Scarton, C., and Specia, L. (2021). The (un)suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, 47(4):861–889.

Alves, A., Miranda, P., Mello, R., and Nascimento, A. (2023). Automatic simplification of legal texts in portuguese using machine learning. In *Legal Knowledge and Information Systems*, pages 281–286. IOS Press.

Baez, A. and Saggion, H. (2023). LSLlama: Fine-tuned LLaMA for lexical simplification. In Štajner, S., Saggio, H., Shardlow, M., and Alva-Manchego, F., editors, *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 102–108, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Chandrasekar, R., Doran, C., and Bangalore, S. (1996a). Motivations and methods for text simplification. In *COLING 1996 volume 2: The 16th international conference on computational linguistics*.

Chandrasekar, R., Doran, C., and Srinivas, B. (1996b). Motivations and methods for text simplification. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

de Menezes, L. C., Paes, A., and Finatto, M. J. B. (2023). Abordagem baseada em aumento de dados para avaliação automática de leiturabilidade. *Domínios de Lingu@gem*, 17:e1721.

Finatto, M. J. B. (2024). Da linguagem simples à acessibilidade textual e terminológica: um percurso com e na linguística. In SILVA, A. H. P.; LAGARES, X. C. M. M. o., editor, *Linguagem simples: para quem? A comunicação cidadã em debate*, pages 86–110. ABRALIN, Campinas.

Heineman, D., Dou, Y., Maddela, M., and Xu, W. (2023). Dancing between success and failure: Edit-level simplification evaluation using SALSA. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3466–3495, Singapore. Association for Computational Linguistics.

Ivanov, V. and Solnyshkina, M. (2020). A method for assessment of text complexity based on knowledge graphs. In *Proceedings of the Linguistic Forum 2020: Language and Artificial Intelligence*, volume 2852 of *CEUR Workshop Proceedings*, Moscow, Russia. CEUR-WS.org.

Kim, T. (2022). Revisiting the practical effectiveness of constituency parse extraction from pre-trained language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5398–5408, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., and Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for army enlisted personnel. Technical Report Research Branch Report 8-75, Chief of Naval Technical Training, Millington, TN, USA.

Leal, S., Duran, M., and Scarton, C. (2024). Nilc-metrix: assessing the complexity of written and spoken language in brazilian portuguese. *Lang Resources & Evaluation*, 58.

Leal, S. E., Duran, M. S., and Aluísio, S. M. (2018). A Nontrivial Sentence Corpus for the Task of Sentence Readability Assessment in Portuguese. In Bender, E. M., Derczynski,

L., and Isabelle, P., editors, *Proceedings of the 27th International Conference on Computational Linguistics*, pages 401–413, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Leal, S. E., Serras, F. R., Finger, M., and Aluísio, S. M. (2026). Complexidade textual e suas tarefas relacionadas. In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, volume 3, book chapter 6. BPLN, 4 edition.

Liu, Y., Fabbri, A., Chen, J., Zhao, Y., Han, S., Joty, S., Liu, P., Radev, D., Wu, C.-S., and Cohan, A. (2024). Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization. In Duh, K., Gomez, H., and Bethard, S., editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4481–4501, Mexico City, Mexico. Association for Computational Linguistics.

Lopes, L. and Pardo, T. (2024). Towards portparser - a highly accurate parsing system for Brazilian Portuguese following the Universal Dependencies framework. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 401–410, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Oliveira, L., Claro, D. B., and Souza, M. (2022). Dptoie: a portuguese open information extraction based on dependency analysis. *Artif. Intell. Rev.*, 56(7):7015–7046.

Queiroz, B., Cavalcante, R., and Claro, D. (2023). Desafios da tarefa de extração de informação aberta: uma abordagem metodológica de um corpus automatizado até o corpus manual. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 388–392, Porto Alegre, RS, Brasil. SBC.

Saggion, H. and Hirst, G. (2017). *Automatic text simplification*, volume 32. Springer.

Scalercio, A., Bertotto, E., Jesus, S., Finatto, M. J., and Paes, A. (2026). Annotation guidelines and challenges for automatic simplification of Portuguese drug leaflets. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 2*, pages 121–127, Salvador, Brazil. Association for Computational Linguistics.

Scalercio, A., Finatto, M., and Paes, A. (2024). Enhancing sentence simplification in Portuguese: Leveraging paraphrases, context, and linguistic features. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Findings of the Association for Computational Linguistics ACL 2024*, pages 15076–15091, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Scalercio, A. M. R. D. A., Souza, E. A. D., Finatto, M. J. B., and Paes, A. (2025). Evaluating LLMs for Portuguese sentence simplification with linguistic insights. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T., editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24452–24477, Vienna, Austria. Association for Computational Linguistics.

Shardlow, M., Alva-Manchego, F., Batista-Navarro, R., Bott, S., Calderon Ramirez, S., Cardon, R., François, T., Hayakawa, A., Horbach, A., Hülsing, A., Ide, Y., Imperial, J. M., Nohejl, A., North, K., Occhipinti, L., Rojas, N. P., Raihan, N., Ranasinghe, T., Salazar, M. S., Štajner, S., Zampieri, M., and Saggion, H. (2024). The BEA 2024 shared task on the multilingual lexical simplification pipeline. In Kochmar, E., Bexte, M., Burstein, J., Horbach, A., Laarmann-Quante, R., Tack, A., Yaneva, V., and Yuan, Z., editors, *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 571–589, Mexico City, Mexico. Association for Computational Linguistics.

Silva, S., Athaydes, A., Mane, B., Claro, D., Souza, M., Neto, F. M., Dantas, L., and Cavalcante, R. (2025). Desafios dos grafos de conhecimento: uma proposta de avaliação de sistemas openie. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 718–722, Porto Alegre, RS, Brasil. SBC.

Vásquez-Rodríguez, L., Shardlow, M., Przybyła, P., and Ananiadou, S. (2021). Investigating text simplification evaluation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 876–882, Online. Association for Computational Linguistics.

Wu, X. and Arase, Y. (2026). An in-depth evaluation of large language models in sentence simplification with error-based human assessment. *ACM Trans. Intell. Syst. Technol.*, 17(4).

Xu, W., Napoles, C., Pavlick, E., Chen, Q., and Callison-Burch, C. (2016). Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*.

Fusão Híbrida para Recuperação de Informação Jurídica Brasileira

Flávio Soriano, Guilherme Evangelista, Celso França, Gisele Pappa, Wagner Meira Jr. e Marcos Gonçalves

[1]Universidade Federal de Minas Gerais, Brasil

{flaviosoriano,guilherme.evangelista,celsofranca}@dcc.ufmg.br
{glpappa,meira,mgoncalv}@dcc.ufmg.br

Abstract. *Brazilian legal retrieval involves heterogeneous search regimes that simultaneously require precise lexical matching and semantic generalization. While hybrid retrieval has become common in general information retrieval, it remains unclear when lexical and semantic signals effectively complement each other in diverse Brazilian legal collections. We investigate this question on JUÁ, a benchmark covering multiple legal retrieval scenarios, including jurisprudence, normative acts, legislative retrieval, and tax-related question answering. We compare BM25, Qwen3 dense retrievers, legal fine-tuned variants, rerankers, and hybrid fusion via Reciprocal Rank Fusion (RRF). Our hypothesis is that lexical-dense fusion can provide greater robustness across heterogeneous legal collections by combining complementary retrieval signals without requiring direct score calibration. Results show that RRF-based systems achieve the best aggregate effectiveness: a legal fine-tuned 4B model leads NDCG@10 and MRR@10, while an 8B model leads MAP@10 and Recall@10. Paired statistical tests show that RRF yields robust gains over BM25 and part of the corresponding dense baselines, although not universally dominating reranking in every individual dataset. Overall, the results suggest that retrieval effectiveness in the legal domain depends strongly on the underlying retrieval regime, and that hybrid lexical-semantic fusion provides a particularly robust configuration for heterogeneous legal search.*

Resumo. *A recuperação de informação jurídica brasileira envolve regimes heterogêneos de busca que exigem simultaneamente correspondência lexical precisa e generalização semântica. Embora a recuperação híbrida seja consolidada na recuperação de informação geral, ainda não está claro quando sinais lexicais e semânticos se complementam efetivamente em coleções jurídicas brasileiras diversas. Investigamos essa questão no benchmark JUÁ, que abrange múltiplos cenários: jurisprudência, atos normativos, recuperação legislativa e perguntas tributárias. Comparamos BM25, recuperadores densos da família Qwen3, variantes ajustadas ao domínio jurídico, reranqueadores e fusão híbrida via Reciprocal Rank Fusion (RRF). Nossa hipótese é que a fusão léxico-densa oferece maior robustez ao combinar sinais complementares sem necessidade de calibração direta de scores. Os resultados mostram que sistemas baseados em RRF alcançam a melhor efetividade agregada: o modelo jurídico ajustado de 4B lidera em NDCG@10 e MRR@10, enquanto o de 8B lidera em MAP@10 e Recall@10. Testes estatísticos pareados indicam*

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 420–432
October 19-22, 2026 ©2026 Association for Computational Linguistics

ganhos robustos do RRF sobre BM25 e parte dos modelos densos, embora não domine todos os reranqueadores individualmente. No geral, a fusão híbrida léxico-semântica revela-se uma configuração particularmente robusta para busca jurídica heterogênea.

1. Introdução

A recuperação de informação jurídica em português brasileiro envolve cenários heterogêneos, que vão da busca jurisprudencial e normativa à recuperação legislativa e perguntas tributárias. Nesses contextos, a relevância de um documento frequentemente depende de correspondência lexical precisa, como termos técnicos, citações normativas e nomes de órgãos, ao mesmo tempo em que consultas em linguagem natural e paráfrases exigem modelos capazes de capturar relações semânticas. Essa tensão torna o domínio jurídico particularmente adequado para investigar a interação entre evidências lexicais, esparsas, e semânticas, densas, em coleções com diferentes regimes de recuperação e necessidades de informação.

Embora estratégias híbridas tenham se consolidado na recuperação de informação geral, ainda há pouca evidência sistemática sobre como sinais lexicais e semânticos se complementam em coleções jurídicas brasileiras heterogêneas. Diferentemente de avaliações centradas em uma única coleção, tarefa ou regime jurídico, este trabalho investiga se uma estratégia simples de fusão por RRF mantém robustez entre regimes jurídicos estruturalmente distintos. Essa questão permanece empiricamente aberta porque a literatura mostra que a eficácia de métodos neurais e híbridos varia conforme domínio, tarefa e disponibilidade de modelos especializados [Thakur et al. 2021, Chen et al. 2022, Louis et al. 2025]. No caso do JUÁ [Pereira et al. 2026], essa questão é particularmente relevante porque os subconjuntos variam não apenas em domínio, mas também em tipo documental, estilo de consulta, granularidade dos julgamentos de relevância e grau de dependência entre correspondência lexical e semântica.

Neste trabalho, investigamos essa questão no benchmark JUÁ [Pereira et al. 2026], uma suíte que reúne diferentes regimes de recuperação jurídica brasileira. Nossa pergunta de pesquisa central é: *em cenários jurídicos heterogêneos, a combinação explícita entre evidência lexical e semântica produz sistemas de recuperação mais robustos do que abordagens puramente densas ou estratégias de reranqueamento?* Para respondê-la, avaliamos BM25, recuperadores densos da família Qwen3-Embedding [Zhang et al. 2025], variantes ajustadas ao domínio jurídico, *pipelines* de reranqueamento e fusão híbrida via *Reciprocal Rank Fusion* (RRF). Nossa hipótese é que a fusão léxico-densa por RRF, ao combinar rankings por posição sem exigir calibração direta de *scores*, pode aproveitar complementaridades entre evidência lexical e semântica, produzindo maior robustez agregada entre coleções jurídicas distintas, mesmo quando não domina todos os subconjuntos individualmente.[1]

Os resultados sustentam essa hipótese no cenário agregado. RRF 4B FT alcança os maiores valores de *NDCG@10* e *MRR@10*, enquanto RRF 8B lidera em *MAP@10* e *Recall@10*. Os testes estatísticos pareados indicam ganhos robustos sobre BM25 e ganhos sobre parte dos recuperadores densos correspondentes, embora a superioridade sobre Rerank não seja uniforme em todos os regimes. Assim, os resultados apontam RRF não

[1] https://github.com/flaviosoriano/fusao-hibrida-recuperacao-informacao-juridica-brasileira

como uma solução universalmente dominante, mas como uma estratégia simples e estável para combinar sinais lexicais e semânticos em um benchmark jurídico heterogêneo.

As principais contribuições deste artigo são:

1. investiga sistematicamente a complementaridade entre sinais lexicais e semânticos em múltiplos regimes de recuperação jurídica brasileira;
2. apresenta uma avaliação comparativa entre recuperação esparsa, densa, reranqueamento e fusão híbrida no benchmark JUÁ;
3. analisa empiricamente em quais cenários evidências lexicais e semânticas atuam de forma complementar;
4. mostra que estratégias híbridas via RRF produzem maior robustez agregada entre coleções heterogêneas, mesmo sem dominância universal em todos os subconjuntos.

2. Trabalhos Relacionados

A recuperação de informação jurídica envolve equilibrar correspondência terminológica precisa com compreensão semântica de consultas em linguagem natural. Métodos lexicais como BM25 [Robertson and Zaragoza 2009] permanecem fortes no domínio jurídico por capturarem termos técnicos, citações normativas e jargões institucionais, enquanto modelos densos, como DPR [Karpukhin et al. 2020] e Sentence-BERT [Reimers and Gurevych 2019], ampliaram a recuperação semântica em cenários de baixa sobreposição lexical. Essa complementaridade motivou adaptações de modelos para *corpora* jurídicos, como LEGAL-BERT [Chalkidis et al. 2020] e abordagens baseadas em BERTimbau [Souza et al. 2020].

Para combinar sinais lexicais e semânticos, a literatura recente explora estratégias de fusão e reranqueamento. O *Reciprocal Rank Fusion* (RRF) [Cormack et al. 2009] combina rankings sem exigir calibração direta de *scores*, enquanto o reranqueamento neural [Nogueira and Cho 2019] reaplica modelos neurais sobre candidatos inicialmente recuperados. Como mostrado em *benchmarks* heterogêneos como BEIR [Thakur et al. 2021], a eficácia varia conforme domínio e tipo de consulta. O benchmark JUÁ [Pereira et al. 2026] formalizou essa diversidade no cenário jurídico brasileiro. Nosso trabalho investiga empiricamente quando a fusão via RRF produz maior robustez do que recuperação densa isolada ou reranqueamento no direito brasileiro.

3. Metodologia

3.1. Preparação dos Dados e Benchmark

Este trabalho avalia métodos de recuperação de informação no benchmark JUÁ, uma suíte pública para recuperação jurídica em português brasileiro [Pereira et al. 2026]. O benchmark reúne coleções com diferentes tipos de documentos e necessidades de informação, incluindo jurisprudência, atos normativos, documentos legislativos e perguntas tributárias. Essa heterogeneidade torna o JUÁ adequado para analisar não apenas o desempenho médio dos sistemas, mas também como diferentes mecanismos de recuperação se comportam em regimes jurídicos distintos.

A tarefa segue a formulação clássica de recuperação ad hoc. Dado um conjunto de consultas Q, um corpus documental D e julgamentos de relevância $\mathcal{R}$, cada sistema deve produzir, para cada consulta, uma lista ordenada de documentos. A qualidade do

ranking é avaliada pela comparação entre os documentos recuperados e os julgamentos de relevância disponíveis. Todos os métodos avaliados neste trabalho utilizam a mesma preparação local dos dados, de modo que as diferenças observadas reflitam os mecanismos de recuperação, e não variações de pré-processamento ou de protocolo experimental.

3.1.1. Coleções avaliadas

A avaliação considera cinco subconjuntos associados ao JUÁ: JUÁ-Juris, juris_tcu, normas_tcu, ulysses_rf e br_taxqa. Os dois primeiros são voltados à recuperação de jurisprudência, enquanto normas_tcu cobre atos normativos do TCU, ulysses_rf representa recuperação legislativa e br_taxqa envolve perguntas em linguagem natural no domínio tributário. Assim, os datasets diferem quanto ao tipo documental, ao formato das consultas e à estrutura dos julgamentos de relevância, permitindo avaliar a robustez dos métodos em cenários jurídicos variados.

Todos os subconjuntos foram convertidos para um formato comum de recuperação, composto por arquivos *corpus.jsonl*, *queries.jsonl* e *qrels_*.tsv*. A preparação local inclui normalização de identificadores, remoção de artefatos simples de HTML, decodificação de entidades, normalização de espaços e conversão dos julgamentos para um formato compatível com o avaliador. Documentos ou consultas sem conteúdo textual útil foram descartados.

3.1.2. Características dos dados

A Tabela 1 resume os subconjuntos avaliados, incluindo tamanho do corpus, número de consultas, quantidade de julgamentos de relevância no conjunto de teste e escala de relevância utilizada. As coleções diferem quanto ao tipo documental, formato das consultas e estrutura dos julgamentos de relevância. Em particular, BR-TaxQA-R enfatiza consultas em linguagem natural, enquanto JurisTCU, NormasTCU e Ulysses-RFCorpus possuem escalas graduadas de relevância.

Tabela 1. Características dos dados preparados localmente para avaliação.

Dataset	Documentos	Consultas	Qrels teste	Escala
JUÁ-Juris	17147	17147	1714	1
JurisTCU	16045	150	2250	0–3
NormasTCU	14469	46	812	0–2
Ulysses-RFCorpus	105669	692	8205	0–2
BR-TaxQA-R	715	715	1211	1–2

3.1.3. Métricas

A avaliação utiliza métricas de ranking no corte 10: *NDCG@10*, *MRR@10*, *MAP@10* e *Recall@10*. *NDCG@10* mede a qualidade posicional considerando graus de relevância; *MRR@10* enfatiza a posição do primeiro documento relevante; *MAP@10* avalia a precisão acumulada nos pontos em que documentos relevantes são recuperados; e *Recall@10* mede a cobertura de documentos relevantes entre os dez primeiros resultados.

Essas métricas permitem distinguir ganhos de ordenação no topo do ranking de ganhos de cobertura, fornecendo uma base adequada para comparar BM25, modelos densos, reranqueamento e fusão híbrida [Kamalloo et al. 2024, Thakur et al. 2021].

3.2. Métodos

Este trabalho propõe e avalia uma estratégia de fusão híbrida para recuperação de informação jurídica brasileira baseada em *Reciprocal Rank Fusion* (RRF). O objetivo é investigar, em um benchmark jurídico heterogêneo, se a combinação entre evidência léxico-sintática e evidência semântica produz rankings mais robustos do que métodos isolados ou pipelines neurais de reranqueamento. A proposta parte da hipótese de que a recuperação jurídica se beneficia da combinação entre sensibilidade à forma textual dos documentos e capacidade de generalização semântica, motivando a combinação de métodos lexicais, como BM25, com modelos densos de recuperação [Nigam et al. 2022].

Dessa forma, organizamos os métodos avaliados em torno de quatro famílias: recuperação lexical, recuperação densa, reranqueamento e fusão por RRF. A proposta central do artigo é avaliar a fusão léxico-densa como uma alternativa simples e robusta para combinar sinais lexicais e semânticos em recuperação jurídica brasileira.

3.2.1. Recuperação lexical e recuperação densa

A recuperação lexical é representada por BM25, utilizado tanto como baseline quanto como uma das entradas da fusão híbrida. BM25 atribui pontuação aos documentos a partir da correspondência entre os termos da consulta e os termos do documento, ponderando frequência lexical, frequência inversa no corpus e normalização por comprimento.

A recuperação densa representa consultas e documentos em um espaço vetorial compartilhado. Cada consulta e cada documento são codificados como vetores, e a ordenação é produzida a partir de uma medida de similaridade entre essas representações. Esse mecanismo busca capturar proximidade semântica mesmo quando há baixa sobreposição lexical.

Neste trabalho, avaliamos dois recuperadores densos gerais, Qwen3-Embedding-4B e Qwen3-Embedding-8B, e duas variantes adaptadas ao domínio jurídico, Qwen3-Embedding-4B FT e Qwen3-Embedding-0.6B FT. No restante do artigo, referimo-nos a esses modelos, respectivamente, como 4B, 8B, 4B FT e 0.6B FT. As variantes FT foram disponibilizadas no ecossistema JUÁ e derivam de ajuste fino de modelos Qwen para recuperação jurídica brasileira, conforme descrito no paper do benchmark [Pereira et al. 2026]. Neste artigo, a adaptação ao domínio não é tratada como substituta da recuperação lexical, mas como uma forma de fortalecer o componente semântico. Assim, uma das questões avaliadas é se a especialização jurídica do modelo denso reduz ou preserva a necessidade de combiná-lo com evidência lexical.

3.2.2. Fusão híbrida por RRF

O componente central da proposta é a fusão de rankings por RRF. A motivação é combinar listas de candidatos produzidas por mecanismos distintos, preservando documentos recuperados por evidência lexical ou semântica. Em um benchmark jurídico heterogêneo,

essa propriedade é desejável porque diferentes coleções podem favorecer diferentes fontes de evidência. Utilizamos RRF porque a fusão é feita com base nas posições dos documentos nos rankings de entrada, e não na combinação direta dos *scores*. Isso é importante porque *scores* de BM25 e similaridades vetoriais não estão naturalmente calibrados na mesma escala. Para um documento d, a pontuação por RRF é definida como $\mathrm{RRF}(d) = \sum_{r \in R} \frac{1}{k + \mathrm{rank}_r(d)}$, em que R é o conjunto de rankings combinados, $\mathrm{rank}_r(d)$ é a posição do documento d no ranking r, e k é uma constante que controla o peso relativo das primeiras posições. Documentos bem posicionados em múltiplos rankings recebem maior pontuação, mas documentos recuperados por apenas um dos mecanismos também podem ser preservados no ranking final.

Em cada variante de fusão, combinamos um ranking lexical produzido por BM25 com um ranking denso produzido por um modelo de embeddings $m \in \mathcal{M}$. Assim, a família de sistemas RRF avaliada neste trabalho pode ser representada por $\mathrm{RRF}_m = \mathrm{RRF}(r_{\mathrm{BM25}}, r_m)$, em que r_{BM25} é o ranking produzido por BM25 e r_m é o ranking produzido pelo recuperador denso m. Essa formulação torna explícito que a proposta não depende de um único modelo denso específico: ela permite avaliar a mesma estratégia de fusão entre evidência léxico-sintática e evidência semântica com diferentes recuperadores densos, incluindo modelos gerais e variantes adaptadas ao domínio jurídico.

3.2.3. Sistemas comparados no benchmark

Para avaliar a proposta, comparamos sistemas que isolam diferentes mecanismos de recuperação: BM25, recuperação densa, recuperação densa adaptada ao domínio, reranqueamento sobre candidatos lexicais e fusão híbrida por RRF. Seja $\mathcal{M}$ o conjunto de modelos densos avaliados: $\mathcal{M} = \{\text{4B}, \text{8B}, \text{4B FT}, \text{0.6B FT}\}$.

Para cada $m \in \mathcal{M}$, avaliamos três usos possíveis do modelo: **Dense** m, isto é, recuperação densa direta; **Rerank** m, isto é, reranqueamento de candidatos recuperados por BM25; e **RRF** m, isto é, fusão entre BM25 e o ranking denso produzido por m. A Tabela 2 resume essas famílias de sistemas.

Tabela 2. Famílias de sistemas avaliadas.

Família	Forma	Mecanismo avaliado
Esparso	BM25	Correspondência léxico-sintática.
Denso	Dense m, para $m \in \mathcal{M}$	Correspondência semântica por embeddings.
Rerank	Rerank m, para $m \in \mathcal{M}$	Reordenação neural de candidatos recuperados por BM25.
Fusão RRF	RRF m, para $m \in \mathcal{M}$	Fusão entre evidência léxico-sintática e semântica.

4. Experimentos

4.1. Configuração Experimental

Os sistemas definidos na Seção 3.2.3 foram executados nos cinco corpora avaliados, usando a preparação local dos dados descrita na Seção 3.1.1. A seguir, descrevemos os parâmetros de execução que diferenciam BM25, Dense, Rerank e RRF.

BM25 foi configurado com $k_1 = 1{,}5$, $b = 0{,}75$ e retorno dos 1000 documentos mais bem classificados por consulta. A tokenização utiliza uma expressão regular

Unicode que preserva números e sequências com hífen, ponto ou barra, o que é relevante para citações normativas e identificadores jurídicos. A recuperação densa utiliza modelos da família Qwen3-Embedding [Zhang et al. 2025], carregados com SentenceTransformer [Wolf et al. 2020], retornando os 1000 documentos mais próximos por consulta.

A fusão RRF combina sempre dois rankings: BM25 e um modelo denso. Utilizamos RRF simples, sem pesos diferenciados, com $k = 60$ e peso $1{,}0$ para cada entrada. O valor $k = 60$ segue a recomendação original de Cormack et al. [Cormack et al. 2009], que o apresentam como uma escolha padrão robusta para RRF. Essa escolha evita introduzir hiperparâmetros adicionais dependentes de dataset e mantém a comparação focada na complementaridade entre o sinal lexical e o sinal semântico, em vez de em uma etapa de calibração ou otimização dos pesos da fusão. A saída contém os 100 documentos mais bem classificados por consulta. Documentos ausentes em um dos rankings não contribuem naquele ranking, e empates são resolvidos pelo identificador do documento.

O reranqueamento, usado como baseline comparativo, é baseado em embeddings bi-encoder com os mesmos modelos de $\mathcal{M}$: para cada consulta, recebe os 1000 candidatos recuperados por BM25, substitui o score original pelo score do reranqueador e retorna os 100 primeiros documentos, usando o rank BM25 original e o identificador do documento como critérios de desempate.

4.2. Resultados Experimentais

Esta seção apresenta os resultados dos sistemas avaliados nos cinco subconjuntos do benchmark e no agregado geral. Primeiro, reportamos o desempenho agregado de todos os sistemas. Em seguida, apresentamos os testes estatísticos pareados em *NDCG@10*. Por fim, resumimos quais sistemas obtiveram os melhores valores por dataset e métrica. As métricas reportadas são *NDCG@10*, *MRR@10*, *MAP@10* e *Recall@10*.

4.2.1. Resultado agregado

Tabela 3. Resultados agregados dos sistemas avaliados. Todas as métricas são calculadas em @10.

Família	Sistema	NDCG	MRR	MAP	Recall
Esparso	BM25	0.4137	0.5223	0.3000	0.4416
Denso	Dense 4B	0.4445	0.5474	0.3329	0.4729
Denso	Dense 8B	0.4507	0.5557	0.3399	0.4722
Denso	Dense 4B FT	0.4562	0.5676	0.3365	0.4845
Denso	Dense 0.6B FT	0.3647	0.4670	0.2591	0.4118
Rerank	Rerank 4B	0.4480	0.5503	0.3352	0.4779
Rerank	Rerank 8B	0.4545	0.5565	0.3431	0.4784
Rerank	Rerank 4B FT	0.4623	0.5740	0.3413	0.4909
Rerank	Rerank 0.6B FT	0.3755	0.4804	0.2678	0.4220
Fusão	RRF 4B	0.4594	0.5701	0.3431	0.4830
Fusão	RRF 8B	0.4690	0.5650	**0.3513**	**0.5009**
Fusão	RRF 4B FT	**0.4744**	**0.5835**	0.3508	0.5006
Fusão	RRF 0.6B FT	0.4326	0.5363	0.3131	0.4708

Os resultados agregados mostram que as melhores configurações observadas pertencem à família de fusão por RRF. RRF 4B FT obtém os maiores valores de *NDCG@10* e *MRR@10*, com 0,4744 e 0,5835, respectivamente. RRF 8B alcança os maiores valores de *MAP@10* e *Recall@10*, com 0,3513 e 0,5009.

BM25 apresenta desempenho agregado inferior aos modelos densos, reranqueadores e fusões. Os modelos densos gerais superam BM25, e Dense 4B FT melhora o desempenho em relação aos modelos densos gerais em *NDCG@10* e *MRR@10*. Entre os reranqueadores, o melhor resultado agregado é obtido por Rerank 4B FT. Ainda assim, as melhores fusões RRF apresentam *scores* agregados superiores aos melhores sistemas densos e de reranqueamento nas métricas principais.

4.2.2. Significância estatística

Aplicamos testes pareados por consulta usando *NDCG@10* como métrica principal. Para cada comparação, calculamos o delta por consulta entre o sistema tratado e a baseline e reportamos a média desses deltas como efeito observado. Intervalos de confiança de 95% foram estimados por bootstrap pareado, enquanto a significância foi avaliada por teste de randomização aproximada unilateral ($tratamento > baseline$), com correção Holm-Bonferroni [Holm 1979] para múltiplas comparações.

Priorizamos o agregado macro estratificado por ser mais alinhado ao objetivo de comparar regimes jurídicos heterogêneos. Nesse agregado, os efeitos são calculados separadamente em cada dataset e combinados por média simples, evitando que coleções maiores dominem a conclusão.

Tabela 4. Testes pareados em *NDCG@10* no agregado macro estratificado.

Comparação	Δ médio	p Holm	Leitura
RRF 4B vs BM25	+0.0457	0.0013	significativo
RRF 8B vs BM25	+0.0553	0.0013	significativo
RRF 4B FT vs BM25	+0.0607	0.0013	significativo
RRF 0.6B FT vs BM25	+0.0189	0.0366	significativo
RRF 4B vs Dense 4B	+0.0149	0.0648	não significativo
RRF 8B vs Dense 8B	+0.0182	0.0184	significativo
RRF 4B FT vs Dense 4B FT	+0.0183	0.0308	significativo
RRF 0.6B FT vs Dense 0.6B FT	+0.0678	0.0013	significativo
RRF 4B vs Rerank 4B	+0.0114	0.1314	não significativo
RRF 8B vs Rerank 8B	+0.0144	0.0620	não significativo
RRF 4B FT vs Rerank 4B FT	+0.0121	0.1314	não significativo
RRF 0.6B FT vs Rerank 0.6B FT	+0.0571	0.0013	significativo
Melhor RRF vs melhor não-RRF	+0.0121	0.1314	não significativo

Os testes estatísticos mostram que todas as variantes RRF superam BM25 de forma estatisticamente significativa. Contra os recuperadores densos correspondentes, RRF apresenta ganho significativo em três das quatro comparações. Já a comparação com reranqueadores é mais heterogênea: apenas RRF 0.6B FT supera significativamente o reranqueador correspondente após correção. Além disso, a comparação entre o melhor RRF e o melhor sistema não-RRF não atinge significância estatística. Portanto, os resultados sustentam ganhos robustos de RRF sobre BM25 e sobre parte dos densos correspondentes, mas não uma dominância estatística universal sobre todos os reranqueadores.

4.2.3. Resultados por dataset

Os melhores sistemas variam entre os regimes avaliados. Dense 8B lidera BR-TaxQA-R; Rerank 0.6B FT lidera JUÁ-Juris; RRF domina JurisTCU e Ulysses-RFCorpus; e Nor-

Tabela 5. Melhores sistemas por dataset e por métrica (calculadas em @10.)

Dataset	Melhor NDCG	Melhor MRR	Melhor MAP	Melhor Recall
BR-TaxQA-R	Dense 8B (0.7879)	Dense 8B (0.8137)	Dense 8B (0.7148)	Dense 8B (0.8612)
JUÁ-Juris	Rerank 0.6B FT (0.2982)	Rerank 0.6B FT (0.2353)	Rerank 0.6B FT (0.2354)	Rerank 0.6B FT (0.5035)
JurisTCU	RRF 4B FT (0.4345)	RRF 4B FT (0.7462)	RRF 4B FT (0.2061)	RRF 4B FT (0.2903)
NormasTCU	RRF 4B FT (0.4092)	Rerank 4B FT (0.5414)	Rerank 4B (0.2916)	Rerank 4B (0.4370)
Ulysses-RFCorpus	RRF 8B (0.6372)	RRF 4B (0.7724)	RRF 4B/8B (0.5127)	RRF 8B (0.6359)

masTCU apresenta resultados divididos entre RRF e reranking. Esses resultados sugerem que a robustez agregada de RRF decorre de desempenho consistente em múltiplos cenários, e não de dominância uniforme em todos os subconjuntos.

5. Análise

Os resultados evidenciam que a recuperação jurídica brasileira não constitui um problema homogêneo, mas um conjunto de regimes de recuperação com demandas distintas de informação. Os achados sugerem que a eficácia relativa entre recuperação lexical, densa e híbrida depende menos do domínio jurídico em si e mais do tipo de evidência predominante em cada coleção. Cenários centrados em correspondência terminológica, citações normativas e jargões institucionais tendem a favorecer mecanismos lexicais como BM25, enquanto consultas formuladas em linguagem natural ou com baixa sobreposição lexical se beneficiam mais da generalização semântica dos modelos densos.

Nesse contexto, o melhor desempenho agregado das estratégias baseadas em *Reciprocal Rank Fusion* (RRF) reflete sua capacidade de combinar sinais complementares sem exigir calibração complexa de *scores*. Ao integrar a precisão léxico-sintática do BM25 com a capacidade de abstração semântica dos recuperadores densos, RRF produz uma configuração particularmente robusta frente à heterogeneidade estrutural do benchmark. A principal exceção ocorre em cenários de perguntas puras em linguagem natural, como no BR-TaxQA-R, em que modelos densos de maior escala (Dense 8B) apresentam compreensão semântica suficiente para reduzir o ganho marginal da evidência lexical. O corpus reduzido, com 715 documentos, também pode contribuir para esse padrão, pois em coleções pequenas o BM25 tende a produzir menos ruído e o ganho marginal da fusão pode diminuir.

Exemplos qualitativos de divergência. Inspecionamos consultas em que BM25 e o recuperador denso atribuíram posições muito distintas ao mesmo documento relevante. Em JUÁ-Juris, na consulta *"A decisão do administrador em não parcelar uma contratação deve ser obrigatoriamente precedida de estudos técnicos que a justifiquem"*: o documento `JURISPRUDENCIA-SELECIONADA-22198` aparece apenas na posição 61 do BM25, mas na posição 8 do modelo denso Qwen3-Embedding-4B FT, sendo promovido para a posição 3 pelo RRF. O trecho relevante discute "contratação por preço global", "ganho de escala" e "contratações parceladas", o que favorece a correspondência semântica em vez da repetição literal da consulta.

O padrão inverso ocorre em Ulysses-RFCorpus, na consulta *"projeto de lei que redefina o ato cooperativo"*: o documento relevante `PLP 271/2005` é recuperado na posição 4 pelo BM25, mas apenas na posição 349 pelo modelo denso Qwen3-Embedding-0.6B FT; o RRF o promove para a primeira posição. Nesse caso, a expressão jurídica

específica "ato cooperativo" favorece a busca lexical, enquanto o modelo denso se desloca para documentos semanticamente próximos, mas inadequados. Esses exemplos ilustram que o RRF preserva evidências complementares: generalização semântica quando há baixa sobreposição lexical e correspondência terminológica quando ela é decisiva.

Geração de candidatos: RRF vs. Reranqueamento. A robustez do RRF em relação aos *pipelines* de reranqueamento avaliados reside na fase de recuperação inicial. O reranqueamento clássico é fundamentalmente limitado pelo *recall* de sua primeira etapa: se o modelo esparso falha em recuperar um documento relevante devido ao descasamento de vocabulário, o reranqueador neural nunca terá a chance de avaliá-lo. O RRF contorna esse gargalo ao compor o *ranking* final a partir de duas distribuições distintas de candidatos. Isso explica por que o reranqueamento permanece forte em subconjuntos onde o BM25 já garante um bom *recall* inicial, mas perde estabilidade no cenário agregado do JUÁ frente à fusão híbrida.

O papel e os limites da adaptação ao domínio. O ajuste fino em *corpora* jurídicos fortalece visivelmente a representação semântica, o que é comprovado pelo RRF 4B FT alcançando os melhores resultados agregados em NDCG@10 e MRR@10. Contudo, os dados provam que a adaptação não elimina a necessidade da fusão léxico-densa. A especialização do modelo reduz o abismo entre a linguagem da consulta e o texto normativo, mas o BM25 continua fornecendo uma evidência ortogonal e estatisticamente significativa. Em suma, o principal mérito do RRF documentado neste trabalho não é uma dominância absoluta em todo e qualquer cenário isolado, mas a sua excepcional estabilidade estrutural frente à severa heterogeneidade da busca jurídica.

6. Conclusão e Trabalhos Futuros

Este trabalho investigou a complementaridade entre recuperação lexical e recuperação densa em recuperação jurídica brasileira heterogênea. Os resultados mostram que variantes baseadas em RRF alcançam as melhores configurações agregadas do benchmark JUÁ, produzindo ganhos robustos sobre BM25 e parte dos recuperadores densos correspondentes. Em particular, RRF 4B FT lidera em *NDCG@10* e *MRR@10*, enquanto RRF 8B obtém os maiores valores de *MAP@10* e *Recall@10*.

Ao mesmo tempo, a análise por dataset evidencia que recuperação jurídica não constitui um problema homogêneo. Consultas em linguagem natural tendem a favorecer recuperadores densos, enquanto tarefas dependentes de correspondência terminológica e citações normativas preservam a importância do sinal lexical. Nesse contexto, o principal mérito do RRF não é uma dominância universal em todos os subconjuntos, mas sua robustez frente a diferentes regimes de recuperação.

Esses resultados sugerem que a eficácia relativa entre recuperação lexical, densa e híbrida depende menos do domínio jurídico e mais da natureza do regime de recuperação predominante em cada coleção. Como trabalhos futuros, pretendemos avaliar reranqueadores mais expressivos, estratégias de fusão ponderada e análises qualitativas capazes de explicar quando evidências lexicais ou semânticas se tornam decisivas para a recuperação.

7. Limitações

Este trabalho apresenta uma avaliação empírica de métodos de recuperação em coleções jurídicas brasileiras associadas ao benchmark JUÁ. Embora os subconjuntos avaliados

cubram jurisprudência, atos normativos, documentos legislativos e perguntas tributárias, eles não representam toda a diversidade de tarefas, instituições, ramos do direito e estilos de consulta presentes na busca jurídica brasileira. Assim, os resultados devem ser interpretados dentro desse escopo experimental.

A comparação com reranqueamento é limitada ao tipo de reranqueador avaliado. Os reranqueadores utilizados baseiam-se em arquiteturas bi-encoder e atuam sobre candidatos recuperados por BM25 — abordagem computacionalmente mais eficiente, mas menos expressiva que métodos cross-encoder, como MonoT5, que podem apresentar comportamento distinto, especialmente em cenários que demandam maior modelagem de interação entre consulta e documento.

Todos os recuperadores densos avaliados pertencem à família Qwen3-Embedding. Modelos multilíngues, como mE5, e arquiteturas orientadas ao português baseadas em BERTimbau não foram avaliados. Portanto, os resultados devem ser interpretados com cautela ao extrapolar para outras arquiteturas de recuperação densa.

A análise estatística também impõe cuidados de interpretação. Os testes principais foram configurados como direcionais, avaliando a hipótese de que o tratamento supera a baseline. Além disso, diferentes formas de agregação respondem a perguntas distintas: o agregado micro atribui maior peso a datasets com mais consultas, enquanto o agregado macro estratificado dá peso comparável aos subconjuntos avaliados.

8. Declaração Ética

Este trabalho avalia métodos de recuperação de informação em documentos jurídicos, mas não propõe substituir a análise de profissionais do Direito. Sistemas de busca jurídica podem recuperar documentos incompletos, enviesados ou fora de contexto, e seus resultados devem ser interpretados como apoio à pesquisa, não como orientação jurídica definitiva. Também é importante considerar que os modelos avaliados podem reproduzir vieses presentes nos dados de treinamento ou nas coleções avaliadas. Assim, qualquer uso prático desses métodos em ambientes jurídicos deve incluir validação humana, transparência sobre suas limitações e cuidado na interpretação dos rankings produzidos.

9. Declaração de Uso de IA Generativa

Os autores utilizaram ferramentas de IA generativa de forma limitada, como apoio editorial e técnico, incluindo revisão gramatical, melhoria de clareza textual, padronização de redação e auxílio inicial na geração de trechos de código. Todo código utilizado nos experimentos foi revisado, adaptado e validado pelos autores. A concepção do estudo, a implementação final dos experimentos, a análise dos resultados, a interpretação dos achados e as conclusões são de responsabilidade integral dos autores. Todas as sugestões geradas por IA foram revisadas criticamente pelos autores antes da submissão.

10. Agradecimentos

Este trabalho foi financiado por CNPq, CAPES, FAPEMIG, FAPESP, AWS, NVIDIA e pelo Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação da Informação (INCT-TILDIAR) (processo nº 408490/2024-1).

Referências

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. (2020). Legal-bert: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904.

Chen, T., Zhang, M., Lu, J., Bendersky, M., and Najork, M. (2022). Out-of-domain semantics to the rescue! zero-shot hybrid retrieval models. In *ECIR 2022*, page 95–110, Berlin, Heidelberg. Springer-Verlag.

Cormack, G. V., Clarke, C. L. A., and Buettcher, S. (2009). Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759. ACM.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Kamalloo, E., Thakur, N., Lassance, C., Ma, X., Yang, J.-H., and Lin, J. (2024). Resources for brewing beir: Reproducible reference models and statistical analyses. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 1431–1440, New York, NY, USA. Association for Computing Machinery.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6769–6781.

Louis, A., van Dijck, G., and Spanakis, G. (2025). Know when to fuse: Investigating non-English hybrid retrieval in the legal domain. In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D., and Schockaert, S., editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4293–4312, Abu Dhabi, UAE. Association for Computational Linguistics.

Nigam, S. K., Goel, N., and Bhattacharya, A. (2022). nigam@coliee-22: Legal case retrieval and entailment using cascading oflexical and semantic-based models. In *New Frontiers in Artificial Intelligence: JSAI-IsAI 2022 Workshop, JURISIN 2022, and JSAI 2022 International Session*, page 96–108.

Nogueira, R. and Cho, K. (2019). Passage re-ranking with bert.

Pereira, J., Fernandes, L., de Brito, E., Lotufo, R., and Bonifacio, L. (2026). Juá – a benchmark for information retrieval in brazilian legal text collections.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992.

Robertson, S. and Zaragoza, H. (2009). The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020*,

pages 403–417. Springer.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., and Gurevych, I. (2021). Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *EMNLP*, pages 38–45.

Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., Xie, P., Yang, A., Liu, D., Lin, J., Huang, F., and Zhou, J. (2025). Qwen3 embedding: Advancing text embedding and reranking through foundation models.

Sector-Specific Financial Sentiment Lexicons for Portuguese: Evidence from Brazilian Stock News

Antônio Artur de Souza[1], **Mateus Binda**[1], **Adriana S. Pagano**[2]

[1] CEPEAD – Centro de Pós-Graduação e Pesquisa em Administração – Faculdade de Ciências Econômicas – Universidade Federal de Minas Gerais (UFMG)
[2]Faculdade de Letras – Universidade Federal de Minas Gerais (UFMG)

artur@face.ufmg.br, tetatrader22334@gmail.com, apagano@ufmg.br

Abstract. *This paper investigates whether sector-specific financial sentiment lexicons improve the analysis of Brazilian stock-market news in Portuguese. We use 19,460 news items about 38 firms in the financial, electric, and manufacturing sectors, published between 2020 and 2025. Starting from OpLexicon, we build sector-specific lexicons through unsupervised expansion based on PMI and conditional word-polarity association. The lexicons are evaluated by lexical heterogeneity and through LSTM price forecasting. Results show sectoral variation and moderate classification performance, but limited predictive gains. Sentiment improves forecasting only in manufacturing, while other sectors do not benefit from its inclusion. The findings indicate that sector-specific lexicons are useful resources, although their predictive value are context dependent.*

1. Introduction

The increasing digitalization of financial markets has substantially expanded the availability of unstructured textual data, including news articles, corporate disclosures, reports, and social media content. In this setting, sentiment analysis has become a relevant strategy for converting qualitative information into quantitative indicators that may support the analysis of market behavior. Prior work has shown that textual sentiment can be associated with prices, returns, and volatility, but the evidence remains mixed and strongly dependent on language, source, time horizon, and modeling choices [Loughran and McDonald 2011, Ranco et al. 2015, Deveikyte et al. 2022, Janková 2023].

For Portuguese, the gap is even more evident. General-purpose lexicons are available, but specialized financial resources remain limited, especially for Brazilian data and sector-specific language use. This limitation matters because the meaning of sentiment-bearing words may change across sectors and market contexts. A term that is positive in one segment may be neutral or even negative in another, which reduces the adequacy of generic lexical resources for financial NLP in Portuguese [Kos et al. 2019, Peres et al. 2019, de Melo 2022].

This paper addresses that gap by investigating whether sector-specific financial sentiment lexicons in Portuguese can better represent Brazilian stock-market news and whether the resulting sentiment signal provides useful downstream information. We

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 433–440
October 19-22, 2026 ©2026 Association for Computational Linguistics

focus on three sectors traded on B3[1]: financial, electric, and manufacturing. From a methodological perspective, we start from a seed lexicon and expand it automatically using an unsupervised procedure based on PMI and conditional polarity association. From an evaluation perspective, we assess the lexicons both intrinsically, through classification-oriented evidence, and extrinsically, through a downstream stock-forecasting task with LSTM models.

Our main contributions are threefold. First, we compile and analyze a Portuguese corpus of 19,460 specialized financial news items related to 38 Brazilian listed companies. Second, we propose sector-specific lexicon expansion for Portuguese financial text and show that the resulting vocabularies display measurable heterogeneity across sectors. Third, we show that the usefulness of the derived sentiment signal is conditional: sentiment is associated with market variables only weakly, and downstream predictive gains are not uniform across sectors. This shifts the discussion away from the binary question of whether sentiment is useful or not, toward the more precise question of when, where, and under what conditions sentiment provides useful information.

2. Related Work

Lexicon-based sentiment analysis remains an important approach when interpretability is desirable and large labeled datasets are not available. However, domain mismatch is a central limitation. Loughran and McDonald [Loughran and McDonald 2011] showed that general-purpose dictionaries often misclassify financial language because many terms common in business discourse do not carry the same polarity they have in everyday language. This observation motivated the construction of domain-specific lexical resources and automatic lexicon-expansion methods for finance and other specialized domains.

In Portuguese, OpLexicon is one of the most widely used lexical resources for opinion mining [Souza et al. 2011]. Even so, it was not originally designed for financial language. Recent work has therefore explored automatic strategies for domain adaptation and lexicon construction in Portuguese, including unsupervised expansion procedures based on distributional and probabilistic signals [de Melo 2021, de Melo 2022, de Sousa and Fernandes 2023]. For the Brazilian financial domain, previous studies have already indicated that lexical methods can extract useful information from news and market-related texts, although performance remains moderate and sensitive to preprocessing and representation choices [Peres et al. 2019, Januário et al. 2022].

The second relevant line of work concerns the relationship between textual sentiment and market variables. Some studies report statistically significant associations between sentiment and returns, prices, or volatility, as well as gains in predictive tasks when text-derived features are incorporated into machine-learning models [Huang and Liu 2020, Hsu et al. 2021, Gao et al. 2022, Zhao et al. 2025]. Others, however, emphasize that these gains are unstable and context-dependent [Ranco et al. 2015, Janková 2023]. This suggests that sentiment should not be treated as a universally informative signal.

Our work lies at the intersection of these two lines of research. We contribute to Portuguese NLP by building sector-specific lexicons for Brazilian financial news and evaluating them not only as lexical resources but also through a downstream forecasting task.

[1]A B3 é a bolsa de valores brasileira e seu nome é uma abreviação para Brasil, Bolsa, Balcão.

3. Corpus and Method

3.1. Corpus construction

The corpus combines financial news retrieved via the GNews API and market data obtained through BRAPI. We collected news about firms listed on B3 and belonging to three sectors: electric, financial, and manufacturing. The sectoral classification follows the institutional structure used by B3 [B3 2025]. After consolidation, auditing, source filtering, and duplicate removal, the final dataset contained 19,460 news items published between November 11, 2020 and November 11, 2025.

The final corpus is distributed across 38 firms and three sectors: 4,286 news items from the electric sector, 11,053 from the financial sector, and 5,112 from manufacturing. The total is not the sum of the sectoral subsets because duplicate cross-sector items were removed from the general corpus. For the downstream forecasting analysis, we selected one representative firm per sector based on market relevance and liquidity: Eletrobras (electric), B3 (financial), and Weg (manufacturing). The retained news sources were specialized financial outlets, namely InfoMoney, Economia UOL, Investing, Valor Investe, and Valor Econômico.

3.2. Seed lexicon and PMI-based title labeling

The starting point for lexicon construction is OpLexicon V3.0 [Souza et al. 2011], a Portuguese opinion lexicon containing lexical items annotated for semantic polarity. Although not financial by design, it offers a suitable seed resource for domain adaptation.

Following Shapiro, Sudhof, and Wilson [Shapiro et al. 2022], we compute PMI over news titles. Titles were chosen because they concentrate the main informational framing of each article and provide a comparable textual unit for the initial labeling stage. Based on the distribution of PMI scores, we assigned titles to three groups using interquartile thresholds: titles at or below the first quartile were labeled as negative, titles at or above the third quartile were labeled as positive, and titles in between were labeled as neutral. This procedure yields balanced positive and negative subsets for subsequent unsupervised expansion.

3.3. Unsupervised sector-specific lexicon expansion

After title labeling, we expanded the seed lexicon using an unsupervised procedure inspired by de Melo [de Melo 2022]. The core idea is to estimate the association between each word and positive or negative classes based on its conditional frequency in previously labeled titles. For each candidate word w_i, we compute positive and negative conditional probabilities and derive a polarity score from their difference. Words with sufficient evidence are then incorporated into the corresponding sector-specific lexicon.

This procedure yields one expanded lexicon for each sector. In practice, it preserves the initial semantic basis provided by the seed lexicon while incorporating vocabulary that is characteristic of sector-specific Brazilian financial news.

3.4. Sentiment scoring

Once the sector-specific lexicons were built, we computed sentiment scores for the news corpus. The sentiment-scoring procedure includes text preprocessing, lexical matching

against the appropriate sector lexicon, and score computation from the balance between positive and negative words, considering negation and modulation cues, following Zhao et al. [Zhao et al. 2025].

The output of this stage is a daily sentiment index for each sector. We then compare these indices with two market variables: same-day closing prices and daily returns. We report Pearson correlations for the full sample and analyze subsamples corresponding to market upturns and downturns, defined as five consecutive trading days with closing prices moving in the same direction.

3.5. Downstream forecasting with LSTM

To test whether sentiment adds downstream predictive information, we trained LSTM models with and without the daily sentiment index. The target is next-day closing price prediction. The architecture contains two stacked LSTM layers with 64 and 32 units, followed by a dense output layer. We used mean squared error as the loss function, the Adam optimizer, and rolling windows of five days as input sequences.

Model performance was assessed through MAE, MSE, and RMSE. The central question is whether adding sentiment to price-based inputs improves forecasting accuracy.

4. Results and Discussion

4.1. PMI distribution and lexical heterogeneity

The PMI distributions suggest a mild positive bias in all sectors, but with different shapes and dispersion. The electric sector shows the highest mean PMI (0.1521), followed by the financial sector (0.1030) and manufacturing (0.0943). This indicates that the informational tone of the news is not distributed uniformly across domains. The title-labeling procedure also yielded balanced positive and negative sets within sectors, supporting the subsequent expansion stage.

Table 1. Descriptive statistics of PMI scores by sector

Statistic	Electric sector	Financial sector	Manufacturing sector
Mean	0.1521	0.1030	0.0943
Standard deviation	0.1798	0.1617	0.1275
Minimum	-0.5098	-0.5819	-0.3811
First quartile	0.0336	0.0010	0.0084
Median	0.1325	0.0963	0.0885
Third quartile	0.2488	0.1890	0.1791
Maximum	0.7792	0.6878	0.5937
Skewness	0.5021	0.3410	0.0737
Kurtosis	0.5466	0.8018	0.1537

The expanded lexicons confirm sectoral heterogeneity. In the electric sector, expansion processed 4,170 labeled titles and incorporated 782 terms, 303 of them positive and 479 negative. In the financial sector, the expanded lexicon reached 1,531 terms, with 606 positive and 925 negative entries. In manufacturing, the expansion produced 849 terms,

356 positive and 493 negative. These differences indicate that sector-specific vocabulary is not merely a matter of word frequency; it also reflects sector-dependent polarity structure.

Intrinsic evaluation results are moderate but informative. For the electric sector, the F_1-score is 0.57 for the positive class and 0.52 for the negative class. In the financial sector, the corresponding values are 0.53 and 0.50. In manufacturing, they are 0.53 and 0.52. The general corpus, in turn, shows values close to the sector-specific ones, which suggests that sector specialization is not primarily justified by a large intrinsic performance jump. Instead, its main relevance lies in capturing domain-specific lexical organization that may matter for interpretation and downstream use.

4.2. Sentiment and market variables

The daily sentiment indices display different statistical profiles across sectors. In the electric sector, the index has mean 3.43 and standard deviation 7.26. In the financial sector, the mean is 8.5 and the standard deviation is 13.67, indicating stronger positive skew and occasional peaks of optimism. In manufacturing, the mean is 3.92 and the standard deviation is 9.34.

The correlations with market variables are statistically detectable in some cases, but they remain weak overall. In the electric sector, sentiment is positively correlated with same-day closing prices (0.0744, $p = 0.0172$), but not with daily returns. In the financial sector, the same-day correlation with closing prices is 0.1123 ($p < 0.01$), while the correlation with daily returns is not significant. Manufacturing exhibits the opposite pattern: sentiment is not significantly associated with same-day closing prices (-0.0367, $p = 0.2311$), but it is positively correlated with daily returns (0.0912, $p < 0.01$).

When we segment the sample by market regime, the electric sector shows a significant positive correlation only during downturns (0.1116, $p < 0.01$). The financial sector shows significant positive coefficients in both upturns (0.1211, $p < 0.01$) and downturns (0.1037, $p < 0.01$). Manufacturing does not show significant regime-specific effects. Taken together, these results indicate that sentiment carries some information, but its relation to market behavior is weak and sector-dependent.

Table 2. Sentiment associations with prices, returns, and market regimes

Parameter	Electric		Financial		Manufacturing	
	Correlation	p-value	Correlation	p-value	Correlation	p-value
Closing price	0.0744^{***}	0.0172	0.1123^{***}	< 0.01	-0.0367	0.2311
Daily returns	0.0488	0.1184	0.0253	0.3800	0.0912^{***}	< 0.01
Upturn phase	0.4135	0.3641	0.1211^{***}	< 0.01	-0.0611	0.1669
Downturn phase	0.1116^{***}	< 0.01	0.1037^{***}	< 0.01	-0.0045	0.9147

Notes: *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

4.3. Downstream evaluation with LSTM

The downstream forecasting task yields the most important practical result of the paper: sentiment does not improve prediction uniformly. In the electric sector, adding sentiment worsens performance substantially. For Eletrobras, MAE increases from 1.25 to 2.72,

MSE from 2.98 to 13.50, and RMSE from 1.72 to 3.67 when sentiment is included. In the financial sector, represented by B3, sentiment also deteriorates the results, though only marginally: MAE changes from 1.5324 to 1.5672, MSE from 4.7055 to 4.9743, and RMSE from 2.1620 to 2.2303.

Manufacturing is the only sector in which sentiment helps. For Weg, the model with sentiment achieves MAE of 2.5473, MSE of 12.6192, and RMSE of 3.5524, compared with 2.7032, 14.0482, and 3.7481 without sentiment. This corresponds to modest improvements of approximately 5.8% in MAE and 4.2% in RMSE.

Table 3. LSTM forecasting results with and without sentiment

Metric	Electric		Financial		Manufacturing	
	With sentiment	Without sentiment	With sentiment	Without sentiment	With sentiment	Without sentiment
MAE (%)	2.72	1.25	1.5672	1.5324	2.5473	2.7032
MSE (%)	13.50	2.98	4.9743	4.7055	12.6192	14.0482
RMSE (%)	3.67	1.72	2.2303	2.1620	3.5524	3.7481
Test sample (days)	241	241	241	241	213	213

The downstream evidence therefore supports restrained interpretation. Sector-specific sentiment lexicons are useful as Portuguese financial NLP resources, but the predictive value of the resulting sentiment index is conditional. In two of the three sectors, adding sentiment introduces noise rather than useful information. Only in manufacturing does the signal appear to complement price history in a consistent way.

5. Conclusion

We presented a study on sector-specific financial sentiment lexicons for Portuguese based on Brazilian stock-market news. Starting from a general-purpose seed lexicon, we built sector-specific lexical resources through unsupervised expansion and evaluated them both intrinsically and extrinsically. The results show that Brazilian financial news exhibits measurable lexical heterogeneity across sectors and that sector-specific lexicons capture part of this structure. However, the usefulness of the resulting sentiment signal is not uniform. Although sentiment is weakly associated with some market variables, downstream forecasting gains appear only in the manufacturing sector.

Overall, the paper contributes to Portuguese NLP by providing evidence that domain and sector adaptation matter for financial sentiment analysis. At the same time, it shows that improved lexical specialization does not automatically translate into robust forecasting benefits. Future work should expand the number of sectors and firms, incorporate manually annotated evaluation data, and compare lexicon-based methods with supervised and transformer-based approaches specialized for Portuguese financial language.

Limitations

This study has some limitations. First, the analysis is restricted to three sectors of the Brazilian stock market and to a specific set of financial news outlets, which may introduce coverage bias. Second, the downstream forecasting analysis relies on a single representative company per sector, which limits firm-level generalization. Third, the lexical approach inherits known limitations of lexicon-based sentiment analysis, including difficulty in handling irony, syntactic scope, semantic shifts, and more subtle contextual cues. Finally,

although the corpus spans five years, the time coverage is bounded by the availability of the news API used for data collection.

Acknowledgments

The authors express their gratitude to the anonymous reviewers for their constructive feedback. Adriana Pagano's research is funded by FAPEMIG and the National Council for Scientific and Technological Development (CNPq, grants 404722/2024-5; 313103/2021-6). This work received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology) and is conducted under the auspices of the National Institute of Science and Technology in Responsible Artificial Intelligence for Computational Linguistics, Information Treatment, and Dissemination (INCT-TILDIAR), funded by CNPq (grant 408490/2024-1).

Ethics Statement

This study uses publicly available news content and publicly available financial market data. No personal, private, or sensitive data were collected. The work is intended as research on language technology and financial text analysis. It should not be interpreted as financial advice or as a recommendation for investment decisions. Given the weak and heterogeneous predictive effects observed in the experiments, the paper explicitly cautions against overinterpreting sentiment signals in applied trading contexts.

Generative AI Usage Disclosure

Generative AI tools were used as auxiliary support in programming, debugging, and language revision tasks. All methodological decisions, data-processing choices, result validation, and final scientific claims were defined, checked, and approved by the authors.

References

B3 (2025). Critério de classificação.

de Melo, T. (2021). Sentiprodbr: Building domain-specific sentiment lexicons for the portuguese language. In *Anais do Simpósio Brasileiro de Banco de Dados*, pages 349–354.

de Melo, T. (2022). Sentilexbr: An automatic methodology of building sentiment lexicons for the portuguese language. *Journal of Information and Data Management*, 13(3):439–449.

de Sousa, T. M. and Fernandes, D. S. A. (2023). Expansão automática de léxico para análise de sentimentos de twitter no domínio do mercado financeiro brasileiro. In *Anais da Escola Regional de Informática de Goiás*.

Deveikyte, J., Geman, H., Piccari, C., and Provetti, A. (2022). A sentiment analysis approach to the prediction of market volatility. *Frontiers in Artificial Intelligence*, 5:1–10.

Gao, Y., Zhao, C., Sun, B., and Zhao, W. (2022). Effects of investor sentiment on stock volatility: new evidences from multi-source data in china's green stock markets. *Financial Innovation*, 8(1):77.

Hsu, Y.-J., Lu, Y.-C., and Yang, J. J. (2021). News sentiment and stock market volatility. *Review of Quantitative Finance and Accounting*, 57(3):1093–1122.

Huang, J.-Y. and Liu, J.-H. (2020). Using social media mining technology to improve stock price forecast accuracy. *Journal of Forecasting*, 39(1):104–116.

Janková, Z. (2023). Critical review of text mining and sentiment analysis for stock market prediction. *Journal of Business Economics and Management*, 24(1):177–198.

Januário, B. A. et al. (2022). Sentiment analysis applied to news from the brazilian stock market. *IEEE Latin America Transactions*, 20(3):512–518.

Kos, S. R., Scarpin, J. E., and Pinto, J. S. d. P. (2019). Proposta de um dicionário de termos de emoção em textos financeiros. *Revista de Ciências da Administração*, 21(55):147–164.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *Journal of Finance*, 66(1):35–65.

Peres, V., Vieira, R., and Bordini, R. (2019). Análises de sentimentos: abordagem lexical de classificação de opinião no contexto mercado financeiro brasileiro. *Anais do Workshop of Artificial Intelligence Applied to Finance*.

Ranco, G. et al. (2015). The effects of twitter sentiment on stock price returns. *PLoS ONE*, 10(9):e0138441.

Shapiro, A. H., Sudhof, M., and Wilson, D. J. (2022). Measuring news sentiment. *Journal of Econometrics*, 228(2):221–243.

Souza, M. et al. (2011). Construction of a portuguese opinion lexicon from multiple resources. In *Proceedings of the Brazilian Symposium in Information and Human Language Technology*.

Zhao, C., Kang, L., Xi, X., Du, S., and Li, J. (2025). Investor sentiment and stock market volatility: exploring the relationship using sentiment analysis of stock bar comments. *Finance Research Open*, 1:100016.

MultiDomainYT-ToxBR: Um Corpus Multidomínio de Comentários Tóxicos do YouTube em Português Brasileiro

Gabriel Z. de Souza[1], **Isabel H. Manssour**[1]

[1]Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), Escola Politécnica
Porto Alegre - RS - Brasil

zurawski.gabriel@edu.pucrs.br, isabel.manssour@pucrs.br

Abstract. *This paper presents MultiDomainYT-ToxBR, a corpus of 5,000 YouTube comments annotated across nine toxicity categories and covering five thematic domains in the Brazilian context. The analysis of annotator agreement highlights inherent challenges, with moderate indices for categories such as racism and xenophobia, whose veiled and context-dependent manifestations hinder consensus. Comparison with ToLD-Br reveals complementary patterns: the new corpus surpasses the previous one in homophobia agreement (α = 0.73 vs. 0.68) and introduces novel categories, such as political and religious intolerance. As validation, BERTweet.BR, after fine-tuning, achieved an F1 of 0.78 in the binary classification.*

Resumo. *Este trabalho apresenta o MultiDomainYT-ToxBR, um corpus de 5.000 comentários do YouTube anotados em nove categorias de toxicidade e distribuídos em cinco domínios temáticos no contexto brasileiro. A análise de concordância entre anotadores evidencia desafios inerentes à tarefa, com índices moderados em categorias como racismo e xenofobia, cujas manifestações veladas e dependentes de contexto dificultam o consenso. A comparação com o ToLD-Br revela padrões complementares: o novo corpus supera o anterior em concordância para homofobia (α = 0,73 vs. 0,68) e introduz categorias inéditas como intolerância política e religiosa. Como validação, o BERTweet.BR, após fine-tuning, atingiu F1 de 0,78 na classificação binária.*

1. Introdução

O crescimento das mídias sociais transformou a comunicação, mas também intensificou o discurso de ódio e o conteúdo tóxico, muitas vezes como mecanismo de engajamento. Evidências indicam que algoritmos de ranqueamento baseados em engajamento amplificam conteúdos emocionalmente carregados e hostis, priorizando a retenção em detrimento da segurança do usuário [Milli et al. 2025]. Esse fenômeno não se restringe ao ambiente digital, pois pesquisas mostram que o ódio disseminado nas redes sociais prediz a ocorrência de crimes violentos contra minorias no mundo real, demonstrando sua influência na propagação de violência offline [Müller and Schwarz 2021].

Casos como a violência contra a minoria muçulmana Rohingya[1] em Mianmar e os motins antimuçulmanos no Sri Lanka[2], ambos associados à amplificação algorítmica de conteúdo hostil no Facebook, ilustram um risco concreto.

[1]https://tinyurl.com/mtrrnxvf
[2]https://tinyurl.com/yetkj4tx

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 441–453
October 19-22, 2026 ©2026 Association for Computational Linguistics

Nesse contexto, a automação da detecção de conteúdo tóxico torna-se uma necessidade. O volume massivo de dados gerados continuamente nas redes sociais inviabiliza a moderação humana, tornando indispensável o desenvolvimento de sistemas de Processamento de Linguagem Natural (PLN) capazes de identificar linguagem nociva em larga escala. Modelos de Linguagem de Grande Escala (LLMs) têm demonstrado potencial promissor nessa tarefa, mas sua eficácia depende da disponibilidade de corpora de treinamento e de avaliação que reflitam a diversidade linguística e cultural de cada contexto.

Este artigo apresenta o MultiDomainYT-ToxBR, um novo corpus de comentários do YouTube em português brasileiro, composto por 5.000 comentários anotados manualmente por múltiplos anotadores em nove categorias de toxicidade, distribuídos em cinco domínios temáticos: política polarizada, futebol, conflitos internacionais, feminismo e direitos das mulheres e conteúdo LGBTQIA+. Este corpus visa suprir a carência de dados do YouTube, complementando os corpora ToLD-Br [Leite et al. 2020] (X, antigo Twitter) e HateBR [Vargas et al. 2022] (Instagram). A especificidade das interações baseadas em vídeos no YouTube justifica a necessidade de um corpus dedicado a compreender suas dinâmicas particulares. As principais contribuições deste trabalho são:

- O MultiDomainYT-ToxBR, um corpus anotado de comentários do YouTube em português, cobrindo múltiplos domínios temáticos;
- Avaliação experimental do corpus, demonstrando sua adequação para classificação binária de toxicidade (F1 de 0,78 com BERTweet.BR, desempenho comparável ao ToLD-Br);
- Análise exploratória da toxicidade no YouTube, evidenciando um perfil mais estrutural e ideológico, com destaque para Intolerância Política;
- Identificação de limitações estruturais da tarefa multilabel, como o impacto do desbalanceamento e da subjetividade na detecção de categorias minoritárias;
- Proposição de uma nova direção de pesquisa, baseada na substituição da toxicidade por tipificação legal, apoiada em corpora anotados por especialistas jurídicos, para reduzir a subjetividade.

2. Trabalhos Relacionados

O ToLD-Br (*Toxic Language Dataset for Brazilian Portuguese*) [Leite et al. 2020] é o maior conjunto de dados para análise de linguagem tóxica em português, com 21.000 tweets do X anotados em sete categorias: não tóxico, LGBTQfobia+, obsceno, insulto, racismo, misoginia e xenofobia. A anotação foi feita por 42 voluntários com perfis demográficos diversos para reduzir vieses. Modelos BERT monolíngues em português alcançaram até 76% de macro-F1 na classificação binária, superando abordagens multilíngues e reforçando a importância de conjuntos de dados específicos para o idioma.

O HateBR [Vargas et al. 2022] é um corpus de 7.000 comentários do Instagram, coletados de contas públicas de políticos brasileiros, para detecção de linguagem ofensiva e discurso de ódio. As anotações, realizadas por indivíduos com alta escolaridade e perfis diversos, abrangem duas camadas: classificação binária (ofensivo/não ofensivo) e nível de ofensividade (leve, moderado, alto), com alta concordância (Kappa de Cohen = 75%).

Mais recentemente, [Trajano et al. 2024] propuseram o OLID-BR, um corpus de 13.538 comentários coletados do X, do YouTube e de *datasets* relacionados (OffComBR [de Pelle and Moreira 2017], HLPHSD [Fortuna et al. 2019], NCCVG [Trajano 2023], ToLD-Br), anotado em um esquema hierárquico de cinco tarefas,

incluindo detecção de alvo e de spans tóxicos. Essa abordagem é inédita para o português brasileiro nesse nível de granularidade.

[Assis et al. 2024] comparam classificadores baseados em BERT com chatbots generativos para detecção de discurso de ódio em português, mostrando que modelos ajustados com poucos exemplos superam os chatbots, embora estes melhorem significativamente com exemplos contextuais. Complementarmente, [Oliveira et al. 2024b] comparam ChatGPT e MariTalk (Sabiá) na mesma tarefa, com o MariTalk demonstrando melhor compreensão do português coloquial e reforçando a eficácia de modelos monolíngues menores. Já [Oliveira et al. 2024a] investigam o LLaMA 3.1 8B com um método iterativo de evolução de *prompts*, obtendo ganhos de F1 em relação ao BERTimbau e precisão comparável à do GPT-4o mini. Por fim, [Junqueira et al. 2024] avaliam o Sabiá-7B em quatro tarefas de PLN, incluindo detecção de ódio e de ironia, com bom desempenho geral em poucos exemplos, mas com limitações na tarefa de pergunta-resposta.

A Tabela 1 sintetiza os trabalhos relacionados. A análise revela uma lacuna: os corpora multilabel disponíveis publicamente para português concentram-se predominantemente no X [Leite et al. 2020] e no Instagram [Vargas et al. 2022]. O OLID-BR [Trajano et al. 2024] incorpora também comentários do YouTube, mas sem estrutura temática por domínio e sem comparação de perfis de toxicidade entre contextos, uma lacuna que o MultiDomainYT-ToxBR busca preencher.

Tabela 1. Síntese dos trabalhos relacionados.

Trabalho	Ano	Modelos Usados	Corporas Usados
[Assis et al. 2024]	2024	BERTimbau, AlBERTina, ChatGPT, BERTweet.BR, MariTalk	HateBR e ToLD-Br
[Oliveira et al. 2024b]	2024	ChatGPT, MariTalk, BERTimbau	ToLD-Br
[Oliveira et al. 2024a]	2024	LLaMA 3.1 8B, BERTimbau, GPT-4o mini, ChatGPT 3.5 Turbo, Sabiá 3	ToLD-Br
[Junqueira et al. 2024]	2024	Sabiá-7B	ABSAPT 2022, ToLD-BR, IDPT 2021, SQUAD v1.1
[Trajano et al. 2024]	2023	BERTimbau, BERT	OffComBR, HLPHSD, NCCVG, ToLD-Br

3. Construção do Corpus

O corpus foi composto por 5.000 comentários em português brasileiro, distribuídos em cinco domínios temáticos propensos à ocorrência de conteúdo controverso e tóxico: Política Polarizada, Futebol, Conflitos Internacionais, Feminismo e Direitos das Mulheres e Conteúdo LGBTQIA+. A escolha desses domínios foi motivada pela representatividade dos debates de alta polarização no contexto brasileiro, o que garante diversidade temática e maior probabilidade de captar diferentes categorias de toxicidade.

Para cada domínio, foram definidos cinco termos de busca, totalizando 25 termos apresentados na Tabela 2. Parte deles foi selecionada manualmente considerando casos de alta repercussão e potencial de toxicidade na mídia brasileira recente, como “xandão e elon musk” e “caso mari ferrer”, enquanto outros foram sugeridos pelo modelo Gemini para complementar e diversificar a cobertura temática de cada domínio.

A coleta seguiu cinco etapas: (1) recuperação de dois vídeos por termo via API do YouTube, totalizando 50 vídeos, filtrados por relevância temática e volume de co-

Tabela 2. Temas e termos de pesquisa associados.

Categorias	Termos de Pesquisa
Política polarizada	"eleições 2024", "pl das fake news", "xandão e elon musk", "lula e bolsonaro", "reforma tributária"
Futebol	"flamengo vs fluminense pós jogo", "corinthians palmeiras torcedor", "gol anulado", "futebol racismo", "torcida organizada briga"
Conflitos internacionais	"guerra ucrânia", "guerra israel e palestina", "crise migratória na Europa", "conclave para escolha de novo papa", "tarifaço trump"
Feminismo e direitos das mulheres	"marieli franco", "violência contra mulher reportagem", "feminismo debate brasil", "caso mari ferrer", "lei maria da penha"
Conteúdo LGBTQIA+	"parada gay", "casamento homoafetivo", "ideologia de gênero", "pronome neutro debate", "educação sexual"

mentários; (2) coleta de todos os comentários desses vídeos; (3) filtragem de idioma com xlm-roberta-base-language-detection[3]; (4) estimativa de toxicidade com Detoxify[4] para amostragem estratificada; (5) seleção de 100 comentários por vídeo (50% sinalizados como potencialmente tóxicos, 50% aleatórios), resultando em 1.000 comentários por domínio.

A anotação considerou nove categorias de toxicidade, baseadas na taxonomia de [Leite et al. 2020] e expandidas para o contexto brasileiro contemporâneo. A Tabela 3 apresenta as categorias e suas definições. Em relação ao ToLD-Br, com seis categorias, o MultiDomainYT-ToxBR introduz três categorias: Intolerância Política, associada à polarização brasileira [Ortellado et al. 2022]; Intolerância Religiosa, impulsionada pelo crescente papel da pauta religiosa no debate público; e Anticapacitismo, comumente expresso como insultos disfarçados em outros contextos.

Tabela 3. categorias de linguagem ofensiva e discriminação.

Categoria	Descrição
Insulto de baixo calão	Agressão verbal com xingamentos, palavrões ou linguagem vulgar, sem conotação discriminatória específica.
Homofobia	Comentários ofensivos baseados em orientação sexual ou identidade de gênero.
Racismo	Comentários hostis ou estereotipados com base em raça, cor da pele ou etnia.
Misoginia/Machismo	Expressões de ódio, desprezo ou inferiorização da mulher enquanto mulher.
Xenofobia	Hostilidade baseada em origem geográfica, nacionalidade ou região, incluindo ataques a migrantes internos.
Violência/Extremismo	Defesa, incentivo ou apologia ao uso de violência física ou psicológica.
Intolerância Religiosa	Ataques ou zombaria com base em crença religiosa.
Intolerância Política	Ataques ou desumanização com base em orientação política ou filiação partidária.
Anticapacitismo	Comentários que zombam ou desumanizam pessoas com deficiência.

A anotação foi realizada em planilhas do Excel e os anotadores foram instruídos a não discutir suas classificações, de modo que interpretações distintas fossem preservadas. Os 5.000 comentários foram divididos em dois grupos independentes de três anotadores cada: o Grupo 1 anotou os primeiros 2.500 comentários e o Grupo 2 anotou os 2.500

[3]https://huggingface.co/papluca/xlm-roberta-base-language-detection

[4]https://pypi.org/project/detoxify/

restantes. A cada categoria foi atribuída uma pontuação de 0 a 3, correspondente ao número de anotadores que a identificaram. Inspirada no ToLD-Br, essa abordagem define uma probabilidade empírica (0/3 a 3/3) por categoria para cada comentário, capturando a ambiguidade inerente à tarefa.

A qualidade das anotações foi avaliada por meio do Kappa de Fleiss e do Alfa de Krippendorff, duas métricas amplamente utilizadas para mensurar concordância em tarefas de anotação com múltiplos avaliadores. A Tabela 4 apresenta as métricas por categoria para cada grupo. Os resultados revelam uma diferença significativa entre os grupos: o Grupo 1 apresenta concordância moderada ($\kappa = 0,493$), enquanto o Grupo 2 atinge apenas o nível justo ($\kappa = 0,372$), segundo os critérios de Landis e Koch [Landis and Koch 1977]. Essa discrepância pode refletir diferenças de perfil e de sensibilidade cultural entre os anotadores, o que reforça a dificuldade inerente à tarefa.

Tabela 4. Kappa de Fleiss por categoria e grupo de anotadores. Valores de Krippendorff's α diferem em menos de 0,0001 em todas as categorias.

Categorias	Grupo 1	Grupo 2
Insulto de baixo calão	0,481	0,262
Homofobia	0,734	0,442
Racismo	0,422	0,332
Misoginia/Machismo	0,532	0,407
Xenofobia	0,329	0,420
Violência/Extremismo	0,364	0,215
Intolerância Religiosa	0,549	0,555
Intolerância Política	0,584	0,499
Anticapacitismo	0,443	0,218
Média Geral	**0,493**	**0,372**

Homofobia apresenta a maior concordância no Grupo 1 ($\kappa = 0,734$), resultado atribuível à presença de marcadores linguísticos explícitos e amplamente reconhecíveis nessa categoria. Intolerância Política e Intolerância Religiosa também demonstram concordância moderada a substancial em ambos os grupos ($\kappa > 0,49$ e $\kappa > 0,54$, respectivamente). No extremo oposto, Violência/Extremismo e Anticapacitismo apresentam os menores índices no Grupo 2 ($\kappa < 0,22$), possivelmente devido à natureza contextual e frequentemente sarcástica dessas manifestações nas discussões online.

Em comparação com o ToLD-Br, cujo α médio é de 0,55, o MultiDomainYT-ToxBR apresenta concordância levemente inferior ($\alpha = 0,493$ no G1), uma diferença de aproximadamente 0,06 pontos, atribuível, em parte, à maior granularidade taxonômica do novo corpus. Homofobia no MultiDomainYT-ToxBR supera a categoria análoga no ToLD-Br ($\alpha = 0,734$ vs. $\alpha = 0,68$), enquanto Racismo e Xenofobia apresentam índices inferiores ($\alpha = 0,422$ e $\alpha = 0,330$ vs. $\alpha = 0,48$ e $\alpha = 0,57$), sugerindo que essas formas de toxicidade se manifestam de modo mais velado no contexto do YouTube do que no X — um achado com implicações diretas para a modelagem.

4. Experimento de Validação

Para verificar a usabilidade do MultiDomainYT-ToxBR como recurso para treinamento supervisionado, foi conduzido um experimento de *fine-tuning* com o modelo

BERTweet.BR [Carneiro et al. 2025], pré-treinado com tweets em português brasileiro, com aproximadamente 110 milhões de parâmetros. A escolha se justifica por seu desempenho superior, reportado em tarefas de detecção de toxicidade em português, quando comparado a alternativas como BERTimbau e ALBERTina [Assis et al. 2024]. Os modelos resultantes desse experimento podem ser encontrados publicamente no Hugging Face[5]. Os *scripts* usados para o *fine-tuning*, bem como os utilizados para a coleta, o tratamento e a análise do corpus, estão disponíveis no GitHub[6].

Os dados foram divididos em 80% treino, 10% validação e 10% teste, por *Hold-Out* estratificado, mantendo a proporção das categorias em cada partição. O tokenizador usado foi o do próprio BERTweet.BR, com limite de 128 tokens. O *fine-tuning* foi conduzido por até 15 épocas com *batches* de 16 exemplos, utilizando o otimizador AdamW com taxa de aprendizado de $2 * 10^{-}5$ e um agendador linear com *warmup* de 500 passos. Foram treinados dois modelos sobre o MultiDomainYT-ToxBR: um para a tarefa de classificação binária (tóxico/não tóxico), com função de ativação *softmax* na inferência, e outro para a tarefa de classificação multilabel (nove categorias), com função sigmoide e limiar de 0,5 para cada categoria. A escolha desses hiperparâmetros baseou-se nos resultados do trabalho de [Assis et al. 2024], que apresentou bons resultados na classificação binária ao ajustar os modelos com o ToLD-Br e o HateBR.

Os resultados da classificação binária são apresentados na Tabela 5. O modelo atingiu F1 de 0,78 no conjunto de teste, com acurácia, precisão e revocação igualmente balanceadas, convergindo já na primeira época. Para fins de referência, o mesmo modelo treinado sobre o HateBR (corpus balanceado artificialmente) atingiu F1 de 0,92, enquanto o treinamento sobre o ToLD-Br resultou em F1 de 0,78. O desempenho inferior do MultiDomainYT-ToxBR em relação ao HateBR é consistente com dois fatores: (i) o desbalanceamento de classes, que reduz a exposição do modelo a exemplos tóxicos durante o treinamento - em contraste com o balanceamento perfeito (50%/50%) do HateBR -; e (ii) a menor concordância entre anotadores observada na Seção 3 ($\kappa = 0,493$ e 0,372, frente ao Kappa de 0,75 do HateBR), que introduz ruído no rótulo-alvo e limita o desempenho máximo alcançável por qualquer classificador supervisionado. Adicionalmente, a maior heterogeneidade temática do corpus (cinco domínios distintos) pode dificultar a generalização, em comparação com o contexto único do HateBR.

Tabela 5. Desempenho do BERTweet.BR na classificação binária sobre o MultiDomainYT-ToxBR.

Corpus	Melhor Época	Acurácia	Precisão	Revocação	F1	Suporte
MultiDomainYT-ToxBR (binário)	1	0,81	0,77	0,79	0,78	500

A tarefa multilabel revelou-se mais desafiadora, conforme apresentado na Tabela 6. O modelo obteve um F1 médio de apenas 0,11, com desempenho concentrado nas duas categorias majoritárias: Insulto de baixo calão (F1 = 0,50) e Intolerância Política (F1 = 0,52). Para as demais sete categorias, o modelo obteve F1 nulo (0,00), falhando em detectá-las no conjunto de teste. Esse resultado era esperado devido ao forte desbalanceamento: categorias como Racismo e Anticapacitismo têm apenas 2 e 3 exemplos no

[5]https://huggingface.co/zurawski

[6]https://github.com/DAVINTLAB/MultiDomainYT-ToxBR

conjunto de teste, tornando inviável o aprendizado de padrões discriminativos por modelos supervisionados.

Tabela 6. Desempenho do BERTweet.BR na classificação multilabel sobre o MultiDomainYT-ToxBR.

Categoria	Precisão	Revocação	F1	Suporte
Insulto de baixo calão	0,63	0,41	0,50	93
Homofobia	0,00	0,00	0,00	15
Racismo	0,00	0,00	0,00	2
Misoginia/Machismo	0,00	0,00	0,00	10
Xenofobia	0,00	0,00	0,00	6
Violência/Extremismo	0,00	0,00	0,00	5
Intolerância Religiosa	0,00	0,00	0,00	7
Intolerância Política	0,69	0,42	0,52	26
Anticapacitismo	0,00	0,00	0,00	3
Média	**0,15**	**0,09**	**0,11**	**167**

5. Análise Exploratória e Discussão

A análise dos corpora MultiDomainYT-ToxBR, ToLD-Br e HateBR revela padrões complementares de toxicidade em português, organizados em quatro eixos descritos a seguir.

5.1. Perfis de Toxicidade por Plataforma

A distribuição das categorias nos três corpora (Figura 1) reflete as características das plataformas de origem e as metodologias de construção. O HateBR apresenta distribuição perfeitamente balanceada (50% tóxico/50% não tóxico), resultado de uma estratégia deliberada de construção que, embora ideal para certas tarefas de treinamento supervisionado, não reflete a prevalência natural do conteúdo tóxico em ambientes reais.

Os demais corpora são desbalanceados e apresentam perfis de toxicidade distintos. No ToLD-Br, a toxicidade é dominada por linguagem explícita: Obsceno (6.652 ocorrências) e Insulto (4.385 ocorrências) concentram a maioria das ocorrências, enquanto Racismo (138) e Xenofobia (151) são marginais. Esse padrão reflete a natureza impulsiva e concisa da comunicação no X, em que a restrição de caracteres favorece ofensas diretas em detrimento de argumentos discriminatórios mais elaborados.

O MultiDomainYT-ToxBR apresenta perfil distinto. Embora Insulto de baixo calão seja a categoria mais frequente (18,5%), o novo corpus se destaca pela maior relevância de toxicidade ideológica e identitária: Intolerância Política surge como segunda categoria mais frequente (5,1%), seguida de Homofobia (3,1%). Esse padrão sugere que, no YouTube, a toxicidade emerge em contextos de debate temático, favorecendo formas mais estruturais e argumentativas de discurso hostil. A ausência de Intolerância Política no ToLD-Br (2019) mostra que corpora antigos podem não capturar dinâmicas sociais recentes, o que torna modelos treinados neles desatualizados.

5.2. Categorias de Baixa Prevalência e Desafios de Detecção

No ToLD-Br, Racismo e Xenofobia representam menos de 1,4% do corpus combinado. No MultiDomainYT-ToxBR, Racismo aparece em apenas 17 comentários (0,34%), Anticapacitismo em 32 (0,64%) e Violência/Extremismo em 57 (1,14%). Essa baixa prevalência não decorre exclusivamente de viés de amostragem. A baixa concordância entre

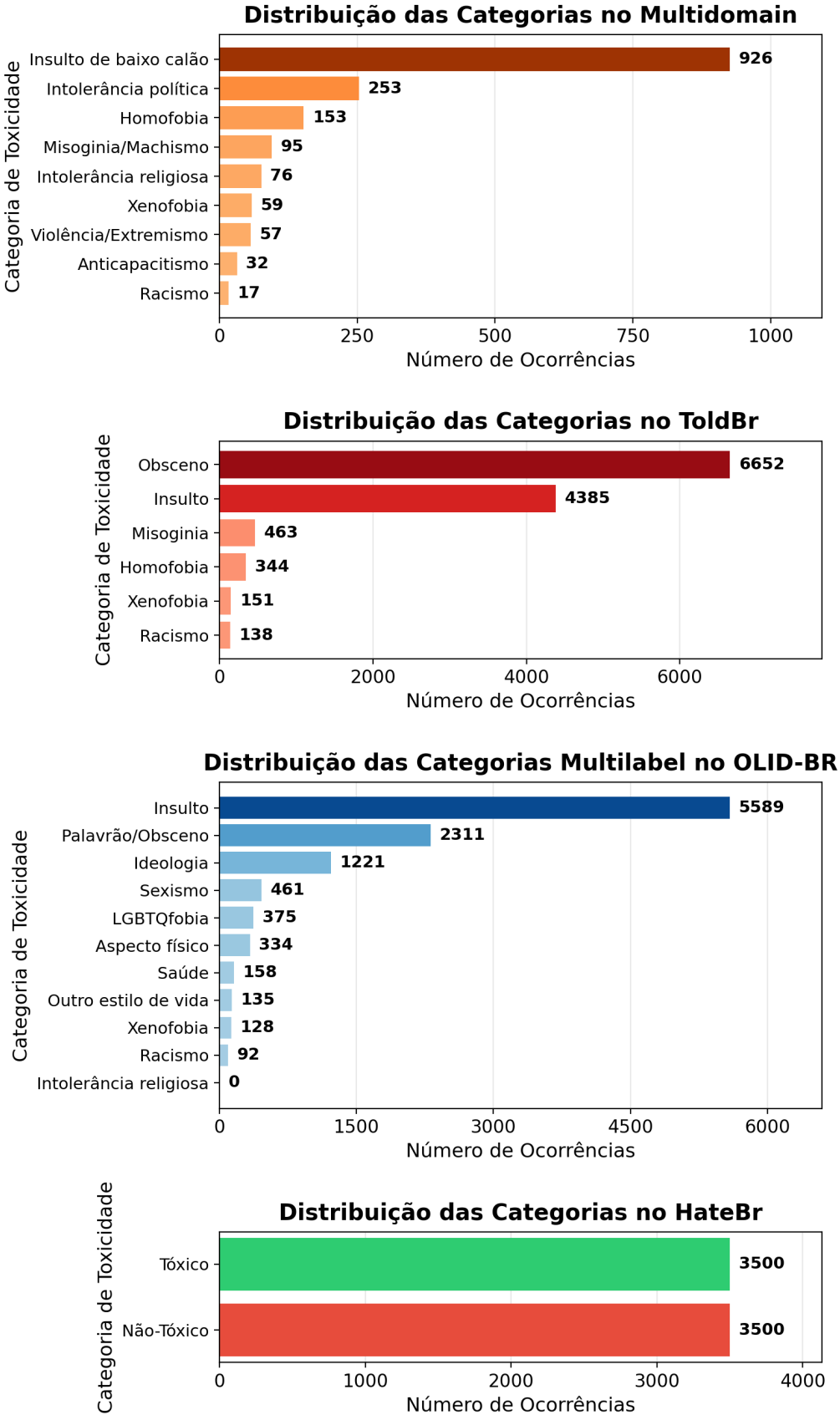


Figura 1. Distribuição das categorias de cada corpus.

anotadores nestas categorias sugere que essas formas de toxicidade se manifestam de modo suficientemente velado ou dependente de contexto para resistir ao reconhecimento consistente por parte de anotadores humanos. Essa combinação de baixa frequência e baixa concordância constitui um obstáculo estrutural à detecção automatizada. Conjuntos de dados com categorias extremamente desbalanceadas levam modelos supervisionados a favorecer as categorias majoritárias e ignorar manifestações tóxicas menos frequentes, mas social e legalmente graves.

5.3. Toxicidade Explícita versus Implícita e Marcadores Linguísticos

A análise de frequência de palavras por categoria[7], descritas nas Tabelas 7 e 8, revela a distinção entre toxicidade explícita e implícita, com implicações para a modelagem. No ToLD-Br, categorias como Homofobia e Misoginia apresentam vocabulário explícito: termos como "viado", "sapatão", "puta" e "piranha" deixam pouca margem para ambiguidade interpretativa. O MultiDomainYT-ToxBR revela um cenário mais complexo. Na categoria Xenofobia, as palavras mais frequentes, "imigrantes", "muçulmanos", "europeus" e "país", são termos neutros que adquirem conotação tóxica apenas em contextos específicos. De modo semelhante, Anticapacitismo apresenta termos explícitos como "retardado", mas também palavras genéricas como "povo", "mundo" e "coisa", indicando que a identificação dessa categoria requer análise contextual que vai além da correspondência por palavras-chave.

Tabela 7. Palavras mais frequentes por categoria de toxicidade - ToldBr.

Homofobia	Obsceno	Insulto	Racismo	Misoginia	Xenofobia
viado (78)	porra (997)	vai (563)	branco (14)	puta (57)	sulista (26)
sapatão (27)	caralho (825)	puta (360)	nego (14)	putinha (56)	carioca (13)
boiola (23)	vai (735)	cara (294)	@user (12)	vai (54)	todo (10)
gay (20)	pqp (710)	porra (272)	vai (10)	piranha (53)	nada (9)
vai (20)	puta (662)	caralho (263)	cara (10)	mulher (30)	gente (8)
caralho (18)	pau (308)	tomar (174)	porra (8)	vagabunda (30)	porra (8)
bicha (17)	tomar (281)	fdp (173)	pau (7)	tudo (23)	vai (8)
cara (15)	vou (273)	gente (161)	ter (7)	caralho (21)	brasil (7)
sapatao (15)	cara (256)	pau (160)	negão (7)	quer (21)	fala (6)
bixa (14)	mano (220)	lixo (154)	agora (6)	fala (18)	ainda (6)

Cada categoria apresenta marcadores linguísticos que atuam como indicadores semânticos. Em Homofobia, termos como "bicha" e "todes", este último frequentemente usado em deboche à linguagem neutra, constituem um vocabulário discriminatório. Em Misoginia, o ToLD-Br apresenta termos sexualizado: "puta", "piranha", "vagabunda"; enquanto no MultiDomainYT-ToxBR emergem termos mais contextuais, como "maria", "penha" e "lei", em referência à Lei Maria da Penha, sugerindo que no YouTube a misoginia se manifesta mais frequentemente em debates sobre direitos das mulheres do que em ofensas diretas. Em Intolerância Política, marcadores como "esquerda", "lula", "direita" e "esquerdista" revelam a forte polarização do contexto brasileiro, com discussões sobre identidade nacional servindo de campo fértil para hostilidade política.

Essa distinção entre toxicidade explícita e implícita tem consequências diretas para o desenvolvimento de sistemas de moderação: abordagens baseadas em listas de

[7]Os termos apresentados nesta seção refletem fielmente o conteúdo dos dados e podem incluir linguagem ofensiva.

Tabela 8. Palavras mais frequentes por categoria de toxicidade - MultiDomainYT-ToxBR.

Insulto	Homofobia	Racismo	Misoginia	Xenofobia	Violência	Into. Relig.	Into. Pol.	Anticap.
vai (103)	mulher (34)	vai (4)	mulher (43)	imigrantes (9)	cara (10)	deus (23)	esquerda (57)	retardado (8)
cara (96)	homem (33)	raça (3)	mulheres (16)	muçulmanos (8)	dessa (7)	religião (21)	lula (25)	povo (4)
merda (74)	vai (19)	macaco (3)	maria (15)	povo (7)	pro (7)	evangélicos (16)	direita (24)	faz (4)
lixo (70)	todes (16)	situação (3)	lei (11)	europeus (7)	vai (7)	sabe (15)	brasil (22)	mundo (4)
mulher (67)	coisa (15)	galo (3)	vida (10)	tudo (7)	inferno (6)	querem (9)	tudo (19)	mulher (4)
brasil (52)	pessoas (15)	racismo (3)	cara (10)	homem (9)	mulher (6)	nada (8)	sempre (15)	retardados (4)
ter (49)	nao (15)	prisão (3)	porque (9)	quer (6)	mano (5)	onde (8)	povo (15)	sabe (4)
faz (46)	deus (15)	egito (2)	penha (9)	serem (6)	mundo (5)	igreja (8)	quer (15)	todes (3)
tudo (45)	gay (14)	agora (2)	feminista (7)	muitos (6)	ter (5)	pessoas (8)	mundo (15)	coisa (3)
todos (44)	pronome (13)	fazer (2)	vida (10)	vai (6)	queimar (4)	vai (7)	esquerdista (15)	ver (3)

palavras-chave ofensivas capturam adequadamente categorias como Insulto e Homofobia, mas falham em categorias cuja toxicidade é construída discursivamente, precisamente aquelas com maior relevância jurídica e social, como Racismo e Xenofobia.

5.4. Implicações Metodológicas

Os resultados obtidos evidenciam limitações que vão além do MultiDomainYT-ToxBR na detecção de toxicidade em português como um todo. Nenhum corpus isolado captura a complexidade da toxicidade em suas múltiplas dimensões: o ToLD-Br, o HateBR e o OLID-BR oferecem cobertura sólida de toxicidade explícita, enquanto o MultiDomainYT-ToxBR contribui com toxicidade estrutural e politizada, mas nenhum dos corpora multilabel apresenta exemplos suficientes de Racismo e Xenofobia. Os resultados também sugerem que a tarefa de classificação baseada em "toxicidade" apresenta uma limitação fundamental: trata-se de um conceito subjetivo, influenciado pelo *background* cultural e pelo nível de tolerância dos anotadores. Assim, os modelos tendem a reproduzir percepções subjetivas de ofensa, em vez de identificar padrões objetivos de conduta ilícita, penalizando contradiscursos legítimos tanto quanto o discurso de ódio que visam combater. Trabalhos futuros devem considerar uma reorientação de paradigma: em vez de classificadores treinados sobre anotações genéricas de toxicidade, reconhecidamente suscetíveis a ruído e viés cultural [Zufall et al. 2022], adotar LLMs instruídos a tipificar textos conforme o Código Penal Brasileiro, distinguindo figuras como injúria racial, apologia ao crime e difamação, validados por corpora com anotações especializadas. [Luo et al. 2023] mostram que esse alinhamento entre PLN e o direito penal não é apenas factível, mas também necessário para que a moderação automatizada de conteúdo tenha respaldo jurídico sólido.

6. Conclusão

Este artigo apresentou o MultiDomainYT-ToxBR, um corpus de 5.000 comentários do YouTube em português, anotados em nove categorias de toxicidade e distribuídos em cinco domínios temáticos. O corpus preenche uma lacuna na literatura ao capturar um perfil de toxicidade mais estrutural, político e ideológico, refletindo a natureza temática dos comentários do YouTube e as dinâmicas sociais brasileiras recentes. A análise comparativa revelou que nenhum corpus isolado captura plenamente a complexidade da toxicidade em português, reforçando a necessidade de avaliações com múltiplos recursos para o desenvolvimento de sistemas robustos. A validação com o BERTweet.BR confirma a adequação do corpus para classificação binária (F1 = 0,78), mas expõe o desbalanceamento das categorias minoritárias como um obstáculo à classificação multilabel, um problema recorrente na área.

7. Limitações

O corpus reflete o debate público brasileiro até 2025, e novos padrões de toxicidade podem não estar representados. Os cinco domínios temáticos, embora diversificados, não cobrem a totalidade dos contextos em que a toxicidade ocorre no YouTube brasileiro. A concordância moderada entre anotadores, especialmente em categorias como Xenofobia e Violência/Extremismo, reflete a dificuldade inerente à tarefa e introduz ruído nos rótulos. Adicionalmente, a divisão dos 5.000 comentários em dois grupos independentes de três anotadores, necessária para reduzir a carga de trabalho voluntário, implica que nenhum anotador avaliou o corpus completo, o que reduz a robustez da anotação global e impede o cálculo de métricas de concordância unificadas entre todos os avaliadores.

Além disso, a etapa de pré-filtragem com a biblioteca Detoxify, utilizada para obter uma amostragem significativa de conteúdo tóxico, introduz um viés sistemático relevante: por ser baseada em um modelo BERT treinado predominantemente em dados em inglês, o Detoxify pode apresentar menor sensibilidade a formas de toxicidade culturalmente específicas do português brasileiro, favorecendo a seleção de conteúdo com marcadores linguísticos mais explícitos e universais em detrimento de manifestações tóxicas mais sutis. Por fim, o experimento se limita ao *fine-tuning* do BERTweet.BR sem técnicas de balanceamento de classes. Arquiteturas alternativas e estratégias, como *oversampling*, estão fora do escopo deste artigo.

8. Considerações Éticas e Uso de IA Generativa

Todos os comentários foram coletados por meio da API oficial do YouTube, em conformidade com seus Termos de Serviço. Os dados são públicos e nenhuma informação privada foi armazenada. No que diz respeito aos anotadores, os seis participantes foram previamente informados sobre a natureza explícita e potencialmente perturbadora do conteúdo a ser classificado, e a participação foi voluntária, podendo ser interrompida a qualquer momento sem qualquer consequência. Quanto ao uso do corpus, este é disponibilizado exclusivamente para fins acadêmicos. Reconhecemos ainda que modelos treinados com dados de toxicidade podem reproduzir vieses presentes nesses dados e recomendamos cautela na implantação em ambientes de produção. Por fim, este artigo contém exemplos de linguagem explícita e ofensiva coletados de redes sociais, apresentados exclusivamente com finalidade analítica e que não representam a opinião dos autores.

Durante a revisão deste artigo, utilizamos o ChatGPT e o Grammarly para aprimorar a clareza e a correção gramatical do texto. Os autores são integralmente responsáveis pelo conteúdo e pela precisão técnica do artigo.

Agradecimentos

Manssour agradece ao CNPq pelo apoio financeiro (processo nº 303208/2023-6).

Referências

Assis, G., Amorim, A., Carvalho, J., de Oliveira, D., Vianna, D., and Paes, A. (2024). Exploring Portuguese hate speech detection in low-resource settings: Lightly tuning encoder models or in-context learning of large models? In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of*

the 16th International Conference on Computational Processing of Portuguese - Vol. 1, pages 301–311, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Carneiro, F., Vianna, D., Carvalho, J., Plastino, A., and Paes, A. (2025). BERTweet.BR: a pre-trained language model for tweets in Portuguese. *Neural Computing and Applications*, 37(6):4363–4385.

de Pelle, R. P. and Moreira, V. P. (2017). Offensive comments in the brazilian web: a dataset and baseline results.

Fortuna, P., Rocha da Silva, J., Soler-Company, J., Wanner, L., and Nunes, S. (2019). A hierarchically-labeled Portuguese hate speech dataset. In Roberts, S. T., Tetreault, J., Prabhakaran, V., and Waseem, Z., editors, *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104, Florence, Italy. Association for Computational Linguistics.

Junqueira, J. d. R., Lopes, , M., C. L. D. S., da Silva, F. L. V., Carvalho, E. A., de Freitas, L. A., and Corrêa, U. B. (2024). Sabiá in action: An investigation of its abilities in aspect-based sentiment analysis, hate speech detection, irony detection, and question-answering. In *IJCNN*, pages 1–8.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Leite, J. A., Silva, D., Bontcheva, K., and Scarton, C. (2020). Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In Wong, K.-F., Knight, K., and Wu, H., editors, *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924, Suzhou, China. Association for Computational Linguistics.

Luo, C., Bhambhoria, R., Dahan, S., and Zhu, X. (2023). Legally enforceable hate speech detection for public forums. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10948–10963, Singapore. Association for Computational Linguistics.

Milli, S., Carroll, M., Wang, Y., Pandey, S., Zhao, S., and Dragan, A. D. (2025). Engagement, user satisfaction, and the amplification of divisive content on social media. *PNAS Nexus*, 4(3):pgaf062.

Müller, K. and Schwarz, C. (2021). Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*, 19(4):2131–2167.

Oliveira, A., Silva, P. H., Santos, V., Moreira, G., Freitas, V. L., and Luz, E. J. (2024a). Toxic text classification in portuguese: Is llama 3.1 8b all you need? In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 57–66, Porto Alegre, RS, Brasil. SBC.

Oliveira, A. d. S., Cecote, T. d. C., Alvarenga, J. P. R., de Souza Freitas, V. L., and da Silva Luz, E. J. (2024b). Toxic speech detection in Portuguese: A comparative study of large language models. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th Internati-*

onal Conference on Computational Processing of Portuguese - Vol. 1, pages 108–116, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Ortellado, P., Ribeiro, M. M., and Zeine, L. (2022). Existe polarização política no brasil? análise das evidências em duas séries de pesquisas de opinião. *Opinião Pública*, 28(1):62–91.

Trajano, D., Bordini, R. H., and Vieira, R. (2024). Olid-br: offensive language identification dataset for brazilian portuguese. *Language Resources and Evaluation*, 58(4):1263–1289.

Trajano, D. d. O. (2023). Detecção de linguagem tóxica aplicada a textos em português. Dissertação (mestrado em ciência da computação), Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, RS, Brasil. Orientador: Rafael Heitor Bordini.

Vargas, F., Carvalho, I., Rodrigues de Góes, F., Pardo, T., and Benevenuto, F. (2022). HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7174–7183, Marseille, France. European Language Resources Association.

Zufall, F., Hamacher, M., Kloppenborg, K., and Zesch, T. (2022). A legal approach to hate speech – operationalizing the EU's legal framework against the expression of hatred as an NLP task. In Aletras, N., Chalkidis, I., Barrett, L., Goanță, C., and Preoțiuc-Pietro, D., editors, *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 53–64, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

An Exploratory Study of Discursive Signals Annotation in User-Generated Content

Rivânia da Paixão de Jesus Carvalho[1], Mateus Araújo Pereira[1], Jackson Wilke da Cruz Souza[1]

[1]Programa de Pós-Graduação em Língua e Cultura
Universidade Federal da Bahia - Salvador/Ba

rivaniacarvalho@bol.com.br {pereiramateus;jacksoncruz}@ufba.br

***Abstract**: In this paper, we present the annotation of Discursive Signals (DSs) in User-Generated Content (UGC), based on a sample of tweets/posts from the financial-market domain annotated according to the Enhanced Rhetorical Structure Theory (eRST) model. The results reveal the presence of recurrent discourse relations in this type of text and their respective DSs, contributing to the expansion of the previously established taxonomy.*

***Resumo:** Nesse artigo, apresentamos a anotação de Sinalizadores Discursivos (SDs) em Conteúdos Gerados por Usuários (UGC), com base em uma amostra de tweets/posts do domínio do mercado financeiro, anotados de acordo com o modelo da Enhanced Rhetorical Structure Theory (eRST). Os resultados revelam a presença de relações discursivas recorrentes nesse tipo de texto e seus respectivos SDs, contribuindo para a expansão da taxonomia previamente estabelecida.*

1. Introduction

Discourse analysis in Natural Language Processing (NLP) relies on theoretical models that derive discourse structure from semantic and syntactic information. One such model is Rhetorical Structure Theory (RST) [Mann and Thompson 1987], which views discourse as organized through coherence relations, allowing the examination of the communicative intentions of authors and possible interpretations derived by readers through the arrangement of Elementary Discourse Units (EDUs). In the context of the processing of Portuguese, discourse-level information plays a central role in tasks such as summarization, opinion mining, argumentation analysis, and dialogue systems [Taboada and Mann 2006].

To account for these phenomena, RST relations are often identified through Discursive Markers (DMs), typically conjunctions and prepositions. Although they can also be realized through syntactic, semantic, lexical, and morpholexical strategies, they are frequently inferred when no explicit markers are present. However, when dealing with less conventional text types, such as User-Generated Content (UGC) [Krumm, Davies, Narayanaswami 2008], RST faces challenges in the consistent identification of discourse relations. In response to these limitations, Zeldes *et al.* (2025) propose Enhanced Rhetorical Structure Theory (eRST), which extends RST by refining segmentation, introducing new relations, and expanding signaling strategies, offering a robust framework for representing discourse across different text types.

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 454–463
October 19-22, 2026 ©2026 Association for Computational Linguistics

Figure 1[1] illustrates an example of discourse annotation based on the eRST framework. The text is segmented into six EDUs, connected through rhetorical relations that organize the informational structure of the discourse. EDUs 1 and 2 are linked by an *Elaboration* relation, in which the second EDU provides additional information about the announced plan. Similarly, EDU 3 is connected to EDUs 4 and 5 through a *List* relation, signaling the enumeration of investment targets. At a higher level, the discourse structure connects the previous segments to EDU 6 through an *Evaluation* relation, expressing the author's interpretation of the reported information.

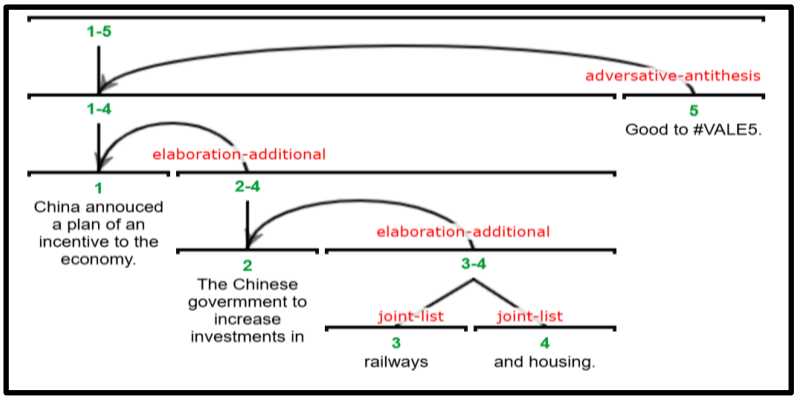


Figure 1 - Discourse Annotation.

The annotated example also presents different DSs that support the interpretation of rhetorical relations in the text. In the *Elaboration* relation between EDUs 2-4 and 3-4, the nominal complement structure *"plan of an incentive"* functions as a *lexical-syntactic* cue. In the *joint-list* relation connecting EDUs 3 and 4, the conjunction "*and*" signals enumeration, while structural parallelism reinforces the interpretation of multiple investment targets. *Evaluation* relation associated with EDU 5 is supported by *evaluative lexical* choices and by the hashtag, which functions as a genre-specific cue frequently observed in financial-domain UGC texts. These elements illustrate how rhetorical relations in UGC may be licensed by lexical, syntactic, structural, and platform-specific cues.

Das and Taboada (2018) argue that DSs must be present whenever a rhetorical relation can be inferred, noting that most relations rely not on explicit DMs, but on structural and linguistic cues. For English, the authors propose a DS taxonomy based on corpora previously annotated by themselves. A DS taxonomy for Brazilian Portuguese was proposed by Dantas *et al*. (2024) developed a DS taxonomy based on the analysis of the CSTNews *corpus* [Cardoso *et al.* 2011], covering Graphical, Semantic, Syntactic, Morphological, and DM cues. This taxonomy was later expanded by Gama and Souza (2024) to better account for UGC, with the introduction of the Genre category, which deals with elements such as emojis and hashtags.

Figure 2 presents the DS taxonomy proposed for Brazilian Portuguese and its extension to UGC texts. The taxonomy organizes signals into six major categories: *Graphical, Semantic, Syntactic, Morphological, Discourse Marker,* and *Genre*. While the first categories capture traditional linguistic signaling strategies, the *Genre* category was introduced to account for platform-specific elements frequently observed in UGC, such

[1] Original tweet: China anunciou um plano de incentivo à economia. O governo chinês pretende aumentar investimentos em ferrovias e moradias. Bom p/ #VALE5

as *hyperlinks, retweets, hashtags*, and *emojis*. This taxonomical organization supports the identification of rhetorical relations beyond prototypical DMs.

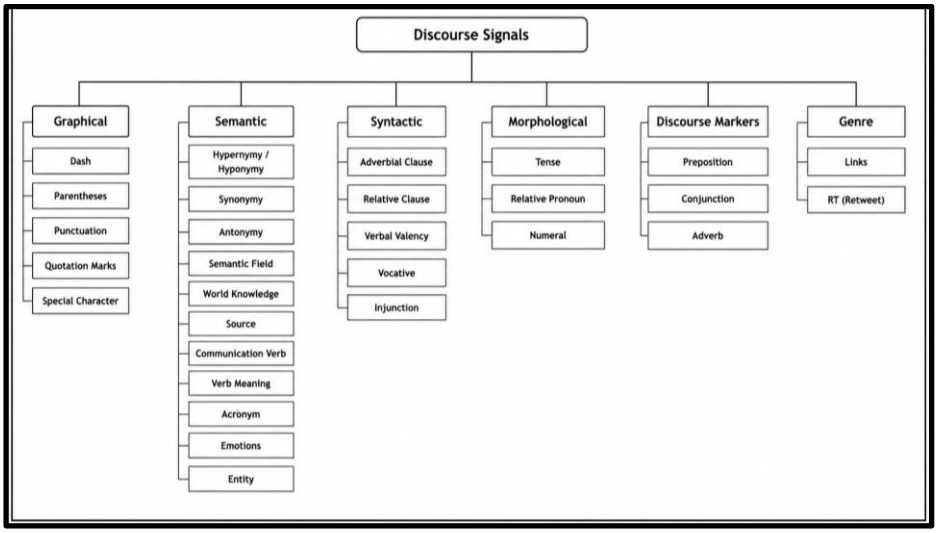


Figure 2 - Discourse Signals translated from Gama and Souza (2024).

Using UGC texts from the financial domain annotated according to the eRST model, researchers conducted both manual and automatic DS identification. In this process, LLMs were employed to detect potential DSs, and their outputs were compared with human annotations, following procedures described in Authors Pereira and Souza (2025). This process resulted in an enhanced *corpus* with an additional annotation layer and contributed to the refinement of the DS taxonomy.

The aim of this study is to identify DSs that can serve as reliable cues for the characterization of eRST relations between EDUs in the financial market tweets/posts domain. To achieve this, we first conducted a manual annotation of EDUs, eRST relations, and their associated DSs in a sample of the *corpus*. Based on this annotation, we describe patterns of association between types of DSs and rhetorical relations in UGC.

A comparative analysis was made between the manual annotation approach and the use of LLMs for the automatic identification of EDUs, eRST relations, and DSs. Subsequently, we examine points of convergence and divergence between the two approaches, as well as their respective limitations. To this end, we compare the model's performance with human annotations and discuss the potential implications for computational discourse modeling.

This ongoing study contributes to this effort by exploring how DSs can be systematically identified, categorized, and partially automated within the eRST framework, with a particular focus on Portuguese tweets/posts from the financial domain. This paper is organized into five sections beyond this Introduction. In Section 2, we present some related works. In Section 3, we outline the methodology employed. Section 4 discusses the results and relevant insights, and in Section 5, we offer concluding remarks.

2. Related work

In recent years, there has been a growing interest in refining LLMs and adapting them to new text genres, particularly UGC, whose informal, fragmented, and platform-dependent nature increases analytical complexity. UGC texts, particularly those published on social

media platforms, exhibit distinctive linguistic features that support the idea that coherence does not depend solely on prototypical DMs but rather on discourse structure [Zeldes *et al.* 2025]. Thus, both the textual properties of the genre and the discursive choices of the interlocutors are shaped by the constraints and affordances of each social media platform.

Launched in 2006, Twitter (now known as "X") has become a significant public forum and has drawn increasing interest from scholars with the aim of understanding user behavior through linguistic manifestations. The texts produced on Twitter platform consist of content shared by individuals or institutions, based on real experiences, evaluations, ideas, comments, and opinions about events, brands, services, or products. According to Santos (2022), these texts can be considered UGC as they are produced and published by users on independent channels of online digital platforms, serving communicative purposes, either individually or in conjunction with other contributions from the same source or other sources.

Di Felippo *et al.* (2021) highlight the unique characteristics of UGC texts on Twitter/X, shaped by platform-specific conventions and limitations, and their implications for automated language processing. Given the informal structure of tweets/posts, the authors identified structural, orthographic and lexical features of UGC texts to define the appropriate unit of analysis.

These characteristics were systematized by Di Felippo *et al.* (2021) into seven dimensions, which include the following phenomena: *code simplification*, such as the addition or substitution of graphic signs like diacritics and hyphens; *abbreviation*, frequently occurring in the form of acronyms; *expression of feelings*, which, most of the time, appear in the form of emoticons and repetition of punctuation; *influence of foreign languages*, especially English; *expression of orality*, which reflects colloquial communication; *metalinguistic elements*, which generally appear at the end of the tweet and are tokenized together with the word or expression they precede; and the *phenomenon of dominance*, which distinguishes DANTEStocks [Di Fellipo *et al.*, 2021] tweets from other tweets/posts.

As eRST aims to offer a more comprehensive representation of rhetorical relations in texts of any genre, as asserted by Zeldes *et al.* (2025), it significantly contributes to the computational modeling of discourse, integrating different levels of linguistic and contextual analysis. This integration has helped to identify real usage patterns in UGC texts. In addition to organizing the relations into a taxonomy adapted to the textual genre, this more comprehensive proposal allows for the characterization of explicit and implicit linguistic signals realized in the text. In line with the studies of [Zeldes *et al*. 2025] and [Di Felippo *et al.* 2021], this research seeks to provide the conditions necessary to capture the complexity, subjectivity, and hybridity typical of UGC texts, based on a taxonomy that enables the investigation and validation of new discursive relations guided by empirical evidence.

In Souza, Cardoso, and Rodrigues (2025) used GPT-4 and LLaMA-4 to identify DMs in news texts, revealing points of convergence and divergence both between the models and in comparison, with human annotation. In the RST relations of Attribution, Elaboration, and Circumstance, LLaMA tends to produce more concise annotations, often reduced to a single central DS such as communication verb *Attribution* or relative clause *Elaboration*. In such cases, GPT-4 offers more detailed and granular descriptions, sometimes excessively, by including additional Graphical, Semantic, or Morphological features. In several instances, both LLMs proposed DSs not included in the existing

taxonomy, such as Subordinate clause, Subordinating conjunction, and Referential element. This indicates potential directions for expanding or revising the typology and highlighting challenges in consistently applying the current labels.

The analysis also revealed certain limitations, including hallucinations and an excessive tendency toward simplification in both GPT-4 and LLaMA, as well as difficulties in identifying DSs in specific EDUs. The results show that the combination of manual and automatic annotation is productive in revising, consolidating, and potentially improving the taxonomy. The study further underscores the need to examine less frequent RST relations, expand the scope to inter-sentential relations, and explore prompt adjustments and hyperparameter tuning. In general, the comparison between models and human annotation highlights both the potential of LLMs for DS annotation and the particularities and limits of their information generation. Among these peculiarities, regarding UGC texts generated on Twitter, in addition to rhetorical relations, we observe DSs with characteristics specific to the genre, as noted by [Zeldes *et al.* 2025].

3. Materials and methods

To perform a comparative analysis between human annotation and the LLMs' performance, in this study, a pilot subset of the DANTEStocks *corpus* [Di Felippo *et al*. 2021] was used, consisting of 60 tweets written in Brazilian Portuguese and collected from the X/twitter platform. Due to the structural variability of the texts, to better account for UGC, the dataset was classified into three categories: Well-structured (WS), Moderately Structured (MS), and Poorly Structured (PS), following the framework proposed by Authors Pereira and Souza (2024).

This classification supports a more precise identification of discourse relations by addressing the distinct phenomena associated with each text type. Table 1 summarizes the criteria adopted for the classification of texts into the WS, MS, and PS categories. This categorization contributes to discourse annotation tasks, since structural variation in UGC directly affects EDU segmentation and rhetorical relation identification.

The subset had previously been manually reviewed and annotated by the Authors Pereira and Souza (2025) using the eRST framework proposed by Amir Zeldes (2025). The manual annotation was carried out by four annotators with academic training in NLP and discourse analysis. Annotation sessions were conducted synchronously and collaboratively, using the rstWeb tool [Zeldes 2016]. Each tweet was annotated by at least two annotators, with disagreements discussed during follow-up sessions.

This is an initial exploratory study whose pilot sample does not yet have quantitative metrics; therefore, the values found are presented in a qualitative and interpretive way, derived directly from the authors' analytical description. However, formal inter-annotator agreement metrics are still under computation. At this stage, the annotation process should be regarded as exploratory, aimed at refining guidelines and identifying ambiguous cases rather than producing a finalized gold standard. The annotator group was guided by the taxonomy and definitions of each type and subtype of DSs proposed by the authors Gama and Souza (2024) and Pereira and Souza (2024)

In addition, as a complementary approach and based on Souza, Cardoso and Rodrigues (2025), we experimented with automatic annotation using a GPT-4 as the LLM. The model was prompted in a few-shot setting to identify: (i) eRST relations, (ii) EDUs, and (iii) associated DSs, based on the taxonomy proposed in earlier work. We

emphasize that the LLM-based annotation is treated as a baseline, not as a replacement for human annotation.

4. Results and Discussion

Based on the DS taxonomy proposed by the Authors Gama and Souza (2024), we identified a range of linguistic, structural, and genre-specific features associated with eRST relations. According to the previous findings of the Authors Pereira and Souza (2025), the most frequent relations observed in this pilot sample were *Attribution-positive, Elaboration-additional, Joint-list, and Organizational-preparation*. This result indicates the prominence of source attribution, informational expansion, and organizational strategies in these types of UGC texts.

In Figure 3, we present examples that compare manual and automatic annotations. These examples are organized according to Pereira and Souza (2024), representing WS, MS, and PS, respectively. Based on these examples, we observe how both approaches (manual and automatic) behave in generating annotation results. At first glance at the examples, variation can be observed in the number of EDUs identified and in discourse segmentation; the manual annotation identified four EDUs, whereas the automatic annotation identified only two. Therefore, they present divergences in the attribution of rhetorical relations and DSs.

In some cases, the same DSs were interpreted as licensing different relations depending on the annotation method, underscoring the inherent ambiguity of DSs in UGC and the sensitivity of annotation outcomes to methodological choices.

A more detailed analysis reveals that these differences are conditioned by the level of *textual structure*. In the WS tweet, manual and automatic annotations show alignment, although the automatic approach simplifies the discourse structure to identify fewer relations and signals. For example, in (3) some eRST relations are preserved across both approaches (e.g., *Organization-preparation* and *Elaboration-additional*); and some relations are identified in a way that partially resembles the manual analysis.

In the MS tweet, we observed partial convergence between both approaches. Differences emerge in the interpretation of relations and in the attribution of DSs, although there is some overlap in the identification of specific signals, such as “RT”, which are identified in both annotations. In the PS tweet, divergences become more pronounced, with substantial variation across both prompted annotation tasks, as well as in the attribution of DSs.

Manual analysis confirms that the eRST framework can be consistently applied to tweets from the financial domain, despite their brevity and informal nature. The recurrence of a small set of relations suggests that beneath the apparent fragmentation of tweets, relatively stable discourse patterns contribute to textual coherence. These patterns are particularly evident in WS tweets, where organizational and elaborative relations are systematically licensed by explicit cues.

Tweets	MANUAL ANNOTATION				AUTOMATIC ANNOTATION			
	RST relationship	Signals			RST relationship	Signals		
		Mark	Type	Subtype		Mark	Type	Subtype
(1) 04/25 The Ilovespa closed down 0.81%, at 51,399 points. Highlights: AEDU3 - 3.72%, MFRG3 - 2.86%, and HGTX - 2.74%.	I. context-circumstance	"04/25"	DM	Adverb	I. organization-preparation	"Highlights"	Lexical	Textual organization
	II. elaboration-additional	"closed down"	Semantics	Verbal sense				
	III. organization-preparation	:	Grafic	Punctuation	II. elaboration-additional	---	---	---
	IV. joint-list	,	Grafic	Punctuation	---	---	---	---
(2) RT @user: Joana is kind of addicted to a game	attribution-positive	RT / @	Text genre	Source	organization-phatic	RT	Pragmatics	Repost
(3) RT @usuário: Look at #PETRA there - Read 'Petrobras debt'...	I. attribution-positive	RT / @	Text genre	Source	organization-phatic	"Look"	---	---
	II. elaboration-additional	"Read"	Semantics	Verbal sense				

Table 1 - Manual and automatic analyses[2].

Regarding discursive signaling, the results demonstrate that eRST relations in UGC are rarely exclusively marked by prototypical DMs. *Attribution-positive* relations, for example, are frequently signaled by platform-specific elements such as retweet markers, user mentions, and explicit references to information sources, although these cues do not necessarily co-occur, especially in poorly structured tweets. *Elaboration-additional* relations are predominantly associated with hyperlinks and named entities (most notably hashtags), which function both as mechanisms of informational expansion and as tools for thematic indexing within the platform.

An important difference between manual and automatic annotation concerns the treatment of DSs and their classification. While manual annotation consistently applies the Type and Subtype distinctions defined in the DS taxonomy, the automatic annotation does not explicitly account for these categories. Nevertheless, the signals identified by LLM provide valuable empirical evidence that can inform future refinements of DSs taxonomy from a computational perspective. In Figure 1, this difference is particularly evident in the automatic annotation to capture signals at a more general level, whereas manual annotation provides a more fine-grained categorization.

Despite the divergences observed between manual and automatic annotation, the results suggest that LLMs approaches can contribute to identifying new possibilities for the classification of signals. As illustrated in Figure 1, the automatic annotation introduces signal types as *lexical* cues, which are not included in DS taxonomies. This evidence suggests that LLMs can reveal patterns that are not captured by current theoretical frameworks.

Analysis of the automatically annotated data further reveals a gradient of discourse organization across the three tweet categories. WS tweets rely primarily on *lexical* and *organizational signals*, such as heading-like expressions (e.g., “Highlights:”), *colon introducers*, *conditional complementizers*, and *quantificational expressions*. These signals systematically license relations such as *organization-preparation*, *joint-list*,

[2] Original tweets: (1) 25/04 O Ibovespa fechou com queda de 0,81%, aos 51.399 pontos. Destaques: AEDU3 (+3,72%), MFRG3 (+2,86%) e HGTX (+2,74%). | (2) RT @usuário: Joana é meio viciada em um jogo. | (3) RT @usuário: Olha a #PETR4 aí. Leia dívida da Petrobras...

elaboration-additional, and *contingency-condition*, yielding relatively explicit nucleus–satellite configurations and a reduced reliance on contextual inference.

MS tweets exhibit a hybrid signaling strategy that combines lexical cues with pragmatic, typographic, and interactional signals (e.g., *quotation marks* and *retweets*). In contrast, PS tweets show a marked decrease in canonical lexical signals and a sharp increase in platform-specific and paralinguistic cues (e.g., *user mentions*, *hashtags*, *emojis*, and *expressive punctuation*). In these cases, discourse coherence is predominantly sustained through organization-phatic and evaluation-comment relations, highlighting the central role of interactional conventions in UGC.

Overall, the comparison between manual and automatic annotation reveals three recurring tendencies: (i) LLMs tend to over-segment EDUs in densely packed or syntactically compressed tweets; (ii) rhetorical relations are often identified correctly at a coarse level but misclassified at finer-grained distinctions; and (iii) discursive signals are frequently detected yet inconsistently mapped onto the predefined Type and Subtype taxonomy. Taking together, these observations suggest that current LLMs implicitly capture discourse-relevant information but lack explicit control over linguistically motivated categories.

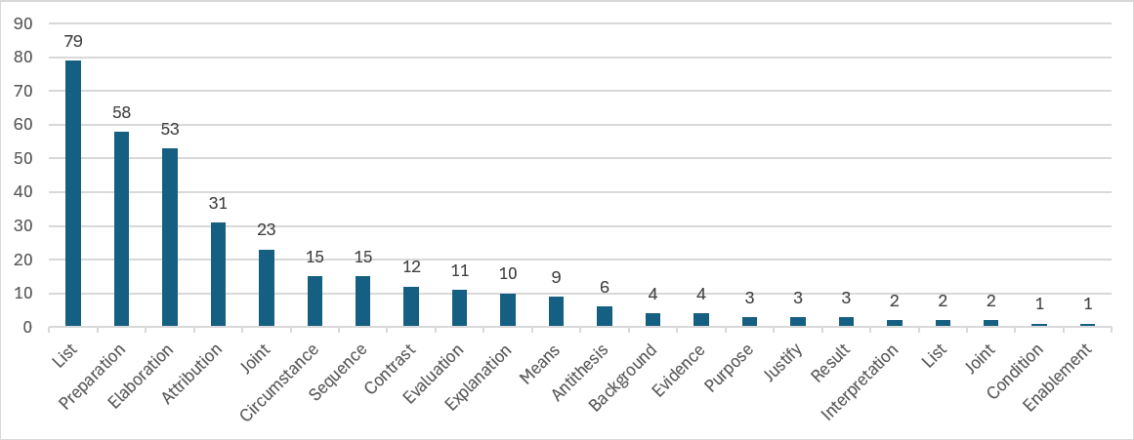


Figure 3 - Quantitative distribution of RST relations on DANTE *corpus*.

5. Final remarks

In this work, the annotation of DSs in Portuguese UGC texts was investigated based on a pilot sample of financial-domain tweets. The application of the eRST framework made it possible to identify recurrent discourse relations and their associated signals, even in a textual genre characterized by brevity, informality, and strong platform constraints. The results indicate that discourse coherence in UGC is sustained by a heterogeneous set of linguistic, structural, and platform-specific cues, whose distribution varies according to the degree of discourse organization of the texts.

From a computational perspective, the exploratory comparison between manual annotation and LLM-based automatic annotation reveals both opportunities and limitations. Although the LLM is often able to identify candidates DSs and rhetorical relations, it lacks systematic control over segmentation decisions and over the mapping of signals to linguistically motivated taxonomies. These findings suggest that LLMs are better viewed, at this stage, as semi-automatic support tools rather than as standalone solutions for discourse annotation.

Furthermore, despite the divergences observed in automatic annotation, the results indicate its potential to suggest new ways of classifying DSs. These results suggest categories that are not represented in existing taxonomies. Rather than being viewed only as a source of inconsistency, it can be considered a valuable source for generating new taxonomic categories for DS classification. This can contribute to future research supporting the expansion and refinement of the classification framework.

The relatively small sample size restricts the generalization of the findings and precludes a robust quantitative evaluation. Therefore, no hyperparameter adjustments were made as this is an initial exploratory study. In addition, the current DSs taxonomy does not yet account for more complex structural properties of discourse, such as hierarchical depth and inter-relational dependencies within the discourse tree.

In future work, we will focus on expanding the *corpus* to include other UGC texts, improving the prompt, refining the DSs taxonomy, and formalizing annotation guidelines. We also plan to scale the annotation process through hybrid human-AI workflows and to introduce standard evaluation metrics for automatic methods. Another promising direction is to compare the distribution of rhetorical relations identified in financial tweets with those observed in other textual genres and domains, to investigate whether this type of UGC presents distinctive discourse organization patterns.

In addition, future research will examine whether providing large language models with texts that have already been segmented into EDUs results in rhetorical analyses that more closely align with human annotations, potentially improving the quality of automatic discourse parsing. This line of research aims to contribute linguistically grounded resources and methodologies that support discourse-aware NLP applications for Portuguese in real-world digital environments.

Acknowledgments

This work was carried out at the Center for Artificial Intelligence of the University of São Paulo (C4AI – http://c4ai.inova.usp.br/), with the support of the São Paulo Research Foundation (FAPESP, Grant No. 2019/07665-4) and IBM Corporation. The project was also supported by the Brazilian Ministry of Science, Technology and Innovation (MCTI), through funding provided under Law No. 8,248 of October 23, 1991, within the scope of the PPI-SOFTEX Program, coordinated by Softex and published as ICT Residency 13 (Official Gazette reference: DOU 01245.010222/2022-44). In addition, this work received financial support from the FAPESB/CNPq Call No. 004/2023 – First Projects Program.

References

Cardoso, P. C., Maziero, E. G., Jorge, M. L. C., Seno, E. M., Di Felippo, A., Rino, L. H. M., Nunes, M. d. G. V., and Pardo, T. A. (2011). Cstnews-a discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese. *In Proceedings of the 3rd RST Brazilian Meeting*, pages 88–105. sn.

Dantas, E., Bárbara, L. de J. S., Pereira, M. A., Gama, N. S., Almeida, T. J. A., Souza, J. W. da C., et al. (2024). Manual de anotação de sinalizadores discursivos em textos jornalísticos. São Carlos: Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. Recuperado de https://repositorio.usp.br/directbitstream/31779c72-b91a-4f9d-8f40-799357711178/3207370.pdf

Das, D. and Taboada, M. (2018). Signalling of coherence relations in discourse, beyond discourse markers. Discourse Processes, 55(8):743–770.

Di Felippo, A., Postali, C., Ceregatto, G., Gazana, L. S., da Silva, E. H., Roman, N. T., and Pardo, T. A. (2021). Descrição preliminar do corpus dantestocks: Diretrizes de segmentação para anotação segundo Universal Dependencies. In Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL), pages 335–343. SBC.

Gama, N., & Souza, J. (2024). *Estudo preliminar sobre sinalizadores discursivos para Conteúdo Gerado por Usuário*. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, (pp. 418-423). Porto Alegre: SBC. doi:10.5753/stil.2024.245060

Mann, W. C. and Thompson, S. A. (1987). Rhetorical structure theory: A framework for the analysis of texts. Technical report.

Narayanaswami, C. (2008). User-generated content. IEEE Pervasive Computing.

Pereira, M., & Souza, J. (2024). Subsídios *Linguísticos para Classificação Automática de Textos de User-Generated Content*. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, (pp. 429-433). Porto Alegre: SBC. doi:10.5753/stil.2024.245132

Pereira, M., & Souza, J. (2025). Anotação Enhanced Rhetorical Structure Theory em textos de User-Generated Content. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, (pp. 713-717). Porto Alegre: SBC. doi:10.5753/stil.2025.37877

Santos, M. L. B. d. (2022). The "so-called" ugc: an updated definition of user-generated content in the age of social media. Online Information Review, 46(1):95–113.

Souza, J. W. da C., Cardoso, P. C. F., & Rodrigues, R. (2025). Taxonomy of discourse signals for RST relations: A study in a news corpus. *Corpus Linguistics: Studies and Applications, 25*(1), e49793. https://doi.org/10.1590/1984-6398202549793

Taboada, M. and Mann, W. C. (2006). Applications of rhetorical structure theory. Discourse studies, 8(4):567–588.

Zeldes, A. (2016). rstweb-a browser-based annotation interface for rhetorical structure theory and discourse relations. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, pages 1–5.

Zeldes, A., Aoyama, T., Liu, Y. J., Peng, S., Das, D., and Gessler, L. (2025). erst: A signaled graph theory of discourse relations and organization. Computational Linguistics, 51(1):23–72.

Do pequeno Asteróide B-612 ao universo: a anotação de O Pequeno Príncipe segundo o modelo *Universal Dependencies*

Magali Sanches Duran, Lucelene Lopes, Maria das Graças Volpe Nunes,

Thiago Alexandre Salgueiro Pardo

Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo

{magali.duran, lucelene}@gmail.com, {gracan, taspardo}@icmc.usp.br

Abstract. *In this paper, we report the annotation process of the iconic book The Little Prince according to the Universal Dependencies international model, as part of an effort to expand the currently available annotated data for Brazilian Portuguese. In particular, we present the annotation protocol and the linguistic decisions involved, as well as a quantitative analysis of the corpus.*

Resumo. *Relatamos, neste artigo, o processo de anotação do icônico livro O Pequeno Príncipe segundo o modelo internacional Universal Dependencies, como parte de um esforço de expansão dos dados anotados atualmente disponíveis para o português do Brasil. Em especial, apresentamos o protocolo de anotação e as decisões linguísticas envolvidas, além de uma análise quantitativa do córpus.*

1. Introdução

A iniciativa internacional *Universal Dependencies* (UD) (de Marneffe et al., 2021), ao agregar mais de 600 contribuidores da Linguística e da Computação de diversas línguas e nacionalidades, vem avançando os estudos da análise de dependências e revolucionando o tratamento sintático na área de Processamento de Línguas Naturais (PLN). Em sua versão atual[1], a UD já conta com a participação de 193 línguas e 353 córpus anotados no mesmo formato[2], o que cria um cenário rico para estudos contrastivos[3] entre línguas e famílias linguísticas, bem como para o desenvolvimento de recursos de PLN multilíngues, como os parsers UDPipe (Straka, 2018) e Stanza (Qi et al., 2020). Graças a esforços recentes, a língua portuguesa passou a ocupar uma posição de destaque, sendo a 5ª língua em termos de número de palavras[4] (segundo dados na listagem online do projeto), contabilizando mais de 1,5 milhão de palavras, sendo que o

[1] https://universaldependencies.org/ (acesso em 23/05/2026)

[2] O formato de anotação da UD é conhecido pela designação CoNLL-U, derivado dos formatos usados nas *shared tasks* da *Conference on Natural Language Learning*.

[3] A comunidade UD organiza um grande evento bianual, o SyntaxFest, dividido em vários sub-eventos que discutem sintaxe empírica, anotação linguística, análise estatística de línguas e PLN. O SyntaxFest está atualmente em sua quinta edição: https://syntaxfest.github.io/syntaxfest27/

[4] Apesar de ser mais usual encontrar esse tipo de contagem em tokens, a página do projeto utiliza o termo *words*, que é o que optamos por mencionar nesta passagem do artigo, para sermos fiéis aos dados oficiais do projeto que estamos citando. A diferenciação entre token e palavra na UD não é uma questão trivial, dado considerar diversos alfabetos: https://universaldependencies.org/u/overview/tokenization.html

ranque é dominado pelo tcheco, com 4,1 milhões de palavras. Em termos de número de córpus, o português também fica em 5º lugar com 7 córpus anotados, sendo que o 1º lugar é ocupado pelo inglês, que conta com 13 córpus anotados.

Para além de sua relevância para os estudos linguísticos, os dados anotados em português têm sido utilizados para diversos fins de PLN, como treinamento de anotadores morfossintáticos (*taggers*) e analisadores sintáticos (*parsers*) (por exemplo, Lopes & Pardo, 2024), embasamento de estudos de níveis linguísticos mais avançados, como relações sintáticas aprimoradas (no inglês, *enhanced relations*) (Duran et al., 2025) e papéis semânticos (Freitas & Pardo, 2025), e desenvolvimento de pesquisas aplicadas variadas, como extração de informação aberta (Oliveira et al., 2023) e caracterização linguística de notícias falsas produzidas por humanos e por máquinas (Andrade et al., 2026).

Visando contribuir para as iniciativas existentes e avançar os estudos para o português do Brasil, relatamos neste artigo o processo de anotação de um córpus construído a partir do livro "O Pequeno Príncipe", de Antoine de Saint-Exupéry. A motivação para essa escolha é o fato de a obra ter sido previamente utilizada por Anchiêta & Pardo (2018) para receber anotação semântica segundo a abordagem *Abstract Meaning Representation* (AMR) (Banarescu et al., 2013) a qual, combinada à anotação UD, pode alavancar muitos estudos na interface entre a sintaxe e a semântica.

As contribuições da presente pesquisa residem não apenas na disponibilização de mais um córpus anotado de acordo com o modelo UD, mas também no fato de que é um córpus de gênero diferente daqueles já anotados em português. Embora a obra seja uma tradução, ela é tomada aqui como legítima representante do gênero literário. Por isso, discutimos seus padrões morfossintáticos e sintáticos como típicos de uma obra literária e acreditamos que decisões de anotação relatadas poderão servir de exemplo para outras anotações do mesmo gênero. É possível, contudo, que futuramente, quando houver outros córpus literários anotados, diferenças entre obras originais e obras traduzidas venham a ser percebidas.

A seguir, apresentamos brevemente um panorama da língua portuguesa no âmbito da UD, assim como as principais características deste modelo. Na Seção 3, relatamos a anotação de córpus que conduzimos, apresentando os dados do córpus, o processo de anotação, as decisões linguísticas tomadas e uma análise quantitativa dos dados anotados. Na Seção 4, tecemos algumas considerações finais.

2. O modelo *Universal Dependencies* e a língua portuguesa

A iniciativa UD (de Marneffe et al., 2021) consiste em um arcabouço para anotação consensual das línguas em nível morfológico, morfossintático e sintático. Além de centenas de atributos morfológicos, são previstas 17 etiquetas morfossintáticas (classes gramaticais; no inglês, *universal part of speech tags* - UPOS tags) e 37 relações de dependência sintática (*dependency relations* - DEPRELs)[5].

[5] As 37 DEPRELs constituem um conjunto fixo de relações de dependência, porém a UD permite que as línguas acrescentem sub-relações para especificá-las. As sub-relações não precisam ser universais, por isso não serão discutidas neste artigo.

Como exemplo de análise UD, reproduzimos na Figura 1 um exemplo de Pagano et al. (2026). As etiquetas morfossintáticas encontram-se abaixo dos tokens, enquanto as relações de dependência estão nos arcos direcionados que conectam os tokens (os atributos morfológicos das palavras não são exibidos).

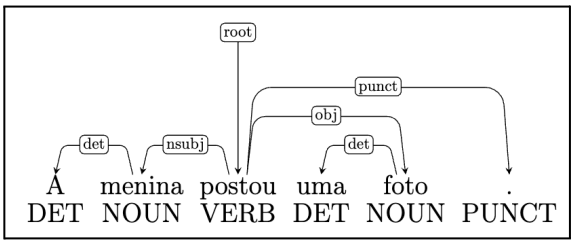


Figura 1. Exemplo de sentença com análise UD (fonte: Pagano et al., 2026)

Pelo que se tem conhecimento, o trabalho pioneiro para o português remonta à pesquisa de Rademaker et al. (2017), que anota um córpus de textos jornalísticos, o Bosque. Desde então, outras iniciativas surgiram, como o PetroGold (Souza et al., 2021), composto por textos acadêmicos do domínio de óleo e gás, e o treebank multigênero Porttinari (Pardo et al., 2021), composto pelos subcórpus Porttinari-base (Duran et al., 2023), com sentenças de notícias jornalísticas da Folha de São Paulo, DANTEStocks, com postagens do Twitter (antes de ser renomeado para X) do domínio financeiro (Di Felippo & Roman, 2026), e PortJur (Lopes et al., 2025), com leis e textos jurídicos (ementas) brasileiros, entre outros. Vários desses esforços foram baseados em manuais e diretrizes de anotação desenvolvidos para o português, como o relatado por Duran et al. (2022)[6].

Até o momento, nenhum córpus de gênero literário em português brasileiro foi disponibilizado no site da UD, lacuna que começará a ser preenchida pelo córpus O Pequeno Príncipe. A seguir, descrevemos o trabalho relativo à anotação desse córpus, que passa a integrar o conjunto de dados anotados com UD para a língua portuguesa, enriquecendo as pesquisas na área.

3. Materiais e Métodos

O córpus O Pequeno Príncipe foi constituído a partir da obra traduzida do francês para o português por Dom Marcos Barbosa, em 1952, e contém 1.527 sentenças e 12.703 tokens. Esse córpus foi criado originalmente por Anchiêta & Pardo (2018) para fazer anotação semântica segundo o modelo AMR (Banerescu et al., 2013) em português.

Aproveitamos, do córpus de Anchieta & Pardo (2018), apenas as sentenças. O alinhamento entre a anotação semântica produzida por Anchiêta & Pardo e a anotação UD, descrita neste artigo, deverá ser feito em trabalhos futuros, pois o formato dessas anotações diferem, o que impede que uma anotação seja simplesmente acrescentada à outra. A anotação AMR não considera todos os tokens, apenas os tokens que expressam conceitos (às vezes um único conceito é expresso por uma combinação de tokens). Já a

[6] A UD evita alterar as diretrizes de anotação a fim de não exigir muita manutenção nos córpus já anotados. O presente trabalho utilizou a versão 2 das Guidelines da UD, a mesma utilizada pelos córpus que constituem a v.2.18 da UD, lançada em 16/05/2026.

anotação UD considera todos os tokens, inclusive os que não são palavras, como é o caso das pontuações.

Normalmente a anotação UD parte de uma anotação automática usando o melhor *parser* disponível no momento, anotação essa que é revisada e corrigida por anotadores humanos, seguindo diretrizes compartilhadas previamente. A anotação automática, por sua vez, pode incluir ou não as fases de segmentação em sentenças e tokenização. Aproveitamos a segmentação feita no córpus original, porém, para a aplicação das diretrizes da UD, a tokenização foi alterada para desmembrar as contrações, por exemplo, "ao" torna-se "a" + "o".

O córpus foi pré-anotado usando o Portparser (Lopes & Pardo, 2024), um *parser* disponível gratuitamente que reporta uma acurácia de 99% na anotação morfossintática e 96% na anotação sintática[7] quando testado no mesmo gênero em que foi treinado, ou seja, no gênero jornalístico. Em seguida, um anotador experiente revisou as sentenças anotadas no ambiente Arborator-Grew (Gibbon et al., 2020), consultando manuais de anotação (Duran et al., 2022). Sentenças que envolviam ambiguidades ou para as quais não havia uma regra clara de anotação foram submetidas à análise de um linguista experiente em UD para decisão sobre a melhor forma de anotar e, se cabível, definição de diretrizes para casos típicos do gênero literário.

Submetemos o córpus revisado ao sistema Verifica-UD (Lopes et al., 2023), uma ferramenta baseada em regras que verifica inconsistências entre a anotação com revisão humana e as regras de anotação em língua portuguesa, produzindo um relatório que orienta a análise e as correções por um linguista. As inconsistências são principalmente detalhes de anotação, como, por exemplo, um verbo anotado com voz passiva (Voice=Pass) cujo sujeito anotado não é o de voz passiva (nsubj:pass). Na sequência, submetemos o córpus ao validador da UD[8], que verifica a aderência da anotação às exigências que a UD faz para todas as línguas antes de incorporar um córpus à sua coleção disponível na web. Por exemplo, todas as expressões fixas têm que ter o Atributo ExtPos anotado, explicitando qual a função morfossintática da expressão como um todo. Esse passo também produziu um relatório de inconsistências que foram resolvidas por um linguista.

3.1. Destaques do processo de anotação

A revisão da anotação automática exige muito conhecimento prévio sobre as diretrizes UD e um olhar atento, pois a tela do Arborator-Grew exibe várias informações e cabe ao anotador/revisor conferir as atribuições de etiquetas e corrigir o que for necessário. A Figura 2 ilustra as etiquetas atribuídas aos tokens de uma sentença (morfossintáticas e sintáticas), além do lema de cada token e atributos dentro das classes gramaticais. O token "*meu*", por exemplo, recebeu a etiqueta morfossintática **DET** (de determinante), a etiqueta sintática **det** (função de determinante), o lema "*meu*" e atributos de tipo de pronome (Prs=pessoal), gênero (Masc=masculino), número (Sing=singular) e pessoa (1=primeira pessoa). Todas essas informações foram geradas automaticamente pelo *parser* e precisaram ser conferidas pelo anotador humano.

[7] https://huggingface.co/spaces/NILC-ICMC-USP/Portparser.v2
[8] https://github.com/UniversalDependencies/tools

Uma grande diferença apresentada pelo gênero literário em comparação a outros gêneros já anotados com UD é a presença de muitas frases interrogativas. O ponto de interrogação ocorre 161 vezes no córpus. Isso representou um desafio quando a pergunta envolvia um pronome interrogativo ou um advérbio, pois o *parser* treinado em outros gêneros nem sempre acertava a atribuição das etiquetas de relações de dependência. A Figura 3 mostra a anotação automática e a versão corrigida de uma sentença do córpus.

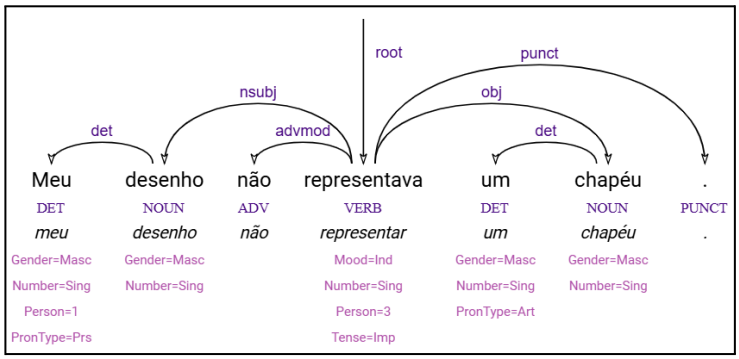


Figura 2. Exemplo de anotação no Arborator-Grew

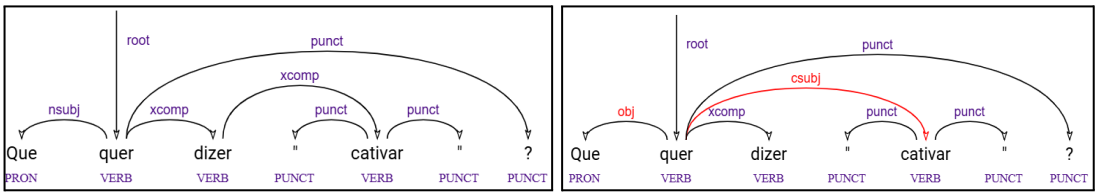


Figura 3. Sentença antes e após a correção

Para realizar essa correção, o anotador precisa perceber que o papel do "*que*" é de objeto direto da locução verbal "*querer dizer*", e "*cativar*" é um sujeito oracional da mesma locução: "*Cativar quer dizer isso*", sendo que "*isso*", no exemplo da interrogativa, é representado pelo pronome "*que*"[9].

Seguindo as diretrizes para português publicadas por Duran et al. (2022), quando se tratava de construção de cópula na interrogativa, havendo dois candidatos a sujeito, anotamos a incógnita como predicativo *(root)*, e o outro constituinte, como sujeito (*nsubj*), como na Figura 4. Na UD não há nenhuma diretriz específica para construções de cópula interrogativas, apenas a diretriz arbitrária para que, numa construção de cópula, o primeiro termo seja anotado como sujeito e o segundo termo como predicativo. Contudo, em construções como "*Onde você está?/Onde está você?*", "*Quando você esteve aqui?*", "*De onde você é?/De onde é você?*", "*Como você está? Como está você?*", é claro que o primeiro termo não poderia ser sujeito, pois não é um nominal. As diretrizes para o português, portanto, estendem essa análise das interrogativas (predicado-sujeito-cópula e predicado-cópula-sujeito) aos pronomes "*que*", "*quem*", "*qual*" e "*quais*": "*Que é isso?*", "*Quem é você?*", "*Qual é o irmão dele?*", "*Quais são as alternativas?*", criando uma recorrência de padrão que beneficia o

[9] "Cativar" foi anotado como VERB e não como NOUN porque, para ser um verbo substantivado, teria que admitir o uso de um determinante, como o artigo "o": *"Que quer dizer [o] "cativar"?".

aprendizado automático sem comprometer a análise linguística[10]. Essa análise, aliás, parece ser compatível com estudos linguísticos funcionalistas, que consideram a incógnita focalizada o elemento mais importante da oração interrogativa e, portanto, o predicativo (v. Fontes, 2012).

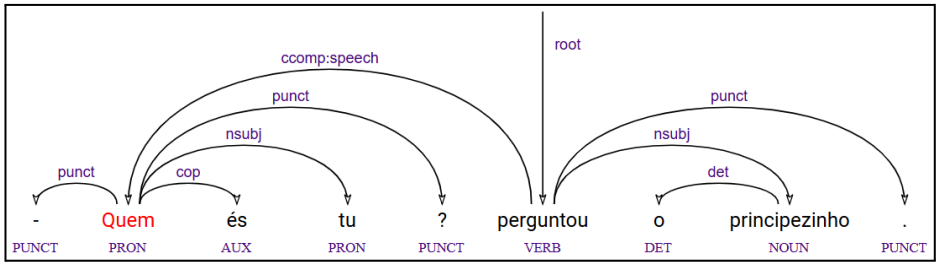


Figura 4. Anotação de construção de cópula interrogativa

O córpus tem o padrão da língua culta vigente nos anos 50 do século XX, com uso de muitos pronomes acusativos e dativos, bem como da segunda pessoa do singular e do plural (*tu* e *vós*), dois fenômenos pouco frequentes no português brasileiro contemporâneo. No córpus, houve 838 ocorrências de verbos na primeira pessoa, 330 na segunda pessoa e 2.210 na terceira pessoa. Em córpus de outros gêneros, como jurídico e jornalístico, é rara a ocorrência de verbos na primeira e na segunda pessoas. A segunda pessoa, em particular, é rara em textos do português brasileiro contemporâneo, uma vez que as formas "*você*"/"*vocês*", mais usuais, constituem uma forma de se referir à segunda pessoa do discurso mas utilizam a flexão de terceira pessoa.

Outras marcas diacrônicas no córpus são encontradas no léxico. A colocação "*fazer medo*" é usada para expressar o que sincronicamente é expresso por "*dar medo*": "*Respondera-me : "Por que é que um chapéu faria medo" ?*".

Encontramos também o uso do clítico junto ao verbo em substituição a um genitivo, o que caracteriza um uso pouco frequente atualmente: "*As únicas montanhas que conhecera eram os três vulcões que* ***lhe*** *davam pelo joelho.*" (Em português brasileiro moderno: "*... que davam/batiam no joelho* ***dele***."). Esse clítico, por ser um pronome objetivo indireto, foi anotado como **iobj** (objeto indireto), apesar de a estrutura argumental do verbo não prever um objeto indireto.

Exceto pelas formas "*lhe*" e "*lhes*", os pronomes acusativos e dativos têm formas iguais e foi preciso atentar para a regência verbal para classificar o pronome como **obj** (case=acc) ou **iobj** (case=dat). Por exemplo, em "*Será uma peça que te prego…*", o pronome "*te*" teve que ser corrigido de **obj** para **iobj** de "*prego*".

A Figura 5 ilustra um problema de tokenização. Se fosse um pronome demonstrativo, "*deste*" deveria ser desmembrado para "*de*" + "*este*". Contudo, nesse córpus ocorreu a forma "*deste*" como a segunda pessoa do singular do passado do verbo "*dar*". Nesse caso, o desmembramento teve que ser desfeito manualmente por não se tratar de uma contração (Figura 5, antes e depois).

[10] Essa análise, contudo, só funciona se o outro termo da cópula for um nominal. Se for um adjetivo ou um advérbio, o pronome é anotado como sujeito. Ex: "*Quem* é mais bonito?", "*Qual* deles é perto daqui?"

Na sentença “*E de novo, sem compreender* ***porque*** *[sic], eu sentia um estranho pesar*”, tokenizamos a forma “*porque*” desmembrando-a em dois tokens: “*por*” + “*que*” uma vez que “*porque*” aí não é uma conjunção, mas sim “*por que*”, um sintagma preposicionado.

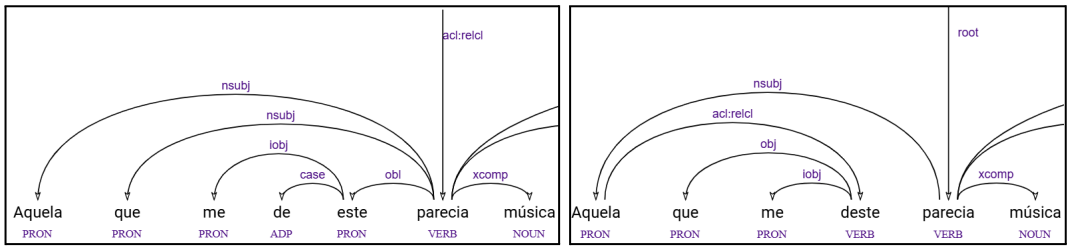


Figura 5. Restauração da forma “deste”, do verbo “dar”

O grande diferencial desse córpus fica por conta da combinação, em uma mesma sentença, de trechos de discurso direto e indireto, o que inexiste nos gêneros já anotados com UD no português. Essa característica acarreta muitas ocorrências de sinais de pontuação raros em outros gêneros, como o ponto de exclamação, o ponto de interrogação e o travessão. Ademais, o uso de inicial maiúscula no interior da sentença (quando há mudança do tipo de discurso), desafia o processamento automático, uma vez que costuma estar correlacionado a início de sentença. O mesmo ocorre com os pontos de interrogação e de exclamação no interior da sentença, pois costumam estar correlacionados a final de sentença.

3.3. Análise quantitativa do córpus anotado

O córpus anotado é composto por 1.527 sentenças. Após os ajustes da tokenização exigidos pela UD, o córpus passou a totalizar 16.802 tokens, representando acréscimo de 4.099 unidades às originais 12.703. O córpus apresenta média de 11 tokens por sentença, considerando palavras, numerais e sinais de pontuação. A Figura 6 apresenta a distribuição do tamanho das sentenças.

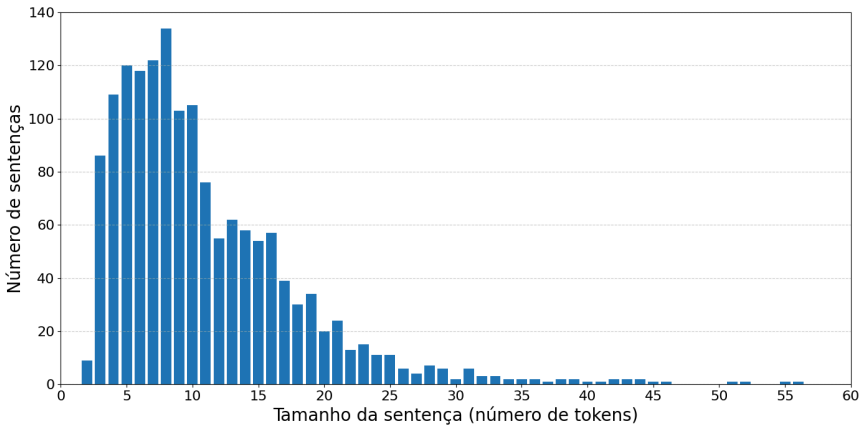


Figura 6. Histograma do número de sentenças por tamanho

Comparado a córpus de outros gêneros, o tamanho médio das sentenças desse córpus pode ser considerado pequeno. No gênero jornalístico, essa média é de 24 tokens (Duran et al., 2022), e no gênero jurídico, de 50 tokens (Lopes et al., 2024).

A UPOS mais frequente do córpus foi PUNCT (3.325 ocorrências), seguida de NOUN (2.585) e VERB (2.215). O fato de o número de verbos (VERB) quase se equiparar ao número de substantivos (NOUN) denota baixa complexidade textual e trabalhos futuros poderão comprovar isso usando métricas (Leal et al., 2023).

As Figuras 7 e 8 apresentam distribuições percentuais comparativas de O Pequeno Príncipe com os três integrantes atuais do córpus multigênero Porttinari: Porttinari-base (textos jornalísticos), PortJur (textos jurídicos) e DANTEStocks (postagens no Twitter do domínio financeiro).

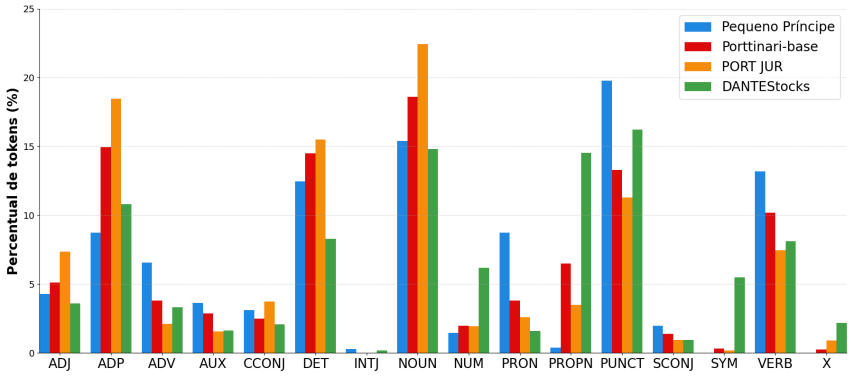


Figura 7. Distribuição das etiquetas UPOS em diferentes córpus

Em relação aos demais gêneros[11], observa-se em O Pequeno Príncipe uma maior frequência de verbos (VERB) e uma menor frequência de substantivos (NOUN) e preposições (ADP). De fato, as nominalizações são uma marca de discursos mais complexos, como o jurídico que, como mostra a Figura 8, é o que apresenta maior percentual de NOUN e ADP. Na Figura 9, a consequência do baixo índice de nominalizações pode ser observada pela frequência da etiqueta de **nmod** (modificador nominal) que, em O Pequeno Príncipe, é a menor de todos os córpus. Observa-se, ainda, frequência relativamente elevada de PRON no córpus, bem como da relação **iobj** (objeto indireto pronominal), o que pode ser explicado pelo uso recorrente de pronomes clíticos dativos, característica da língua padrão vigente à época da tradução. Esse uso, conforme destacam Morais & Salles (2024), está praticamente extinto no português contemporâneo para as formas "*lhe*" e "*lhes*", mas ainda resiste para as formas de primeira e segunda pessoas.

A maior frequência de pronomes de O Pequeno Príncipe em relação aos demais córpus pode também estar relacionada ao fato de se tratar de uma obra traduzida do francês. Como o francês não é uma língua que admite a elipse do sujeito[12], todo sujeito

[11] As barras do gráfico são percentuais de cada etiqueta em relação ao total de tokens de cada córpus. Isso significa que as diferenças de tamanho dos córpus pouco interferem na comparação.
[12] Característica de línguas chamadas *pro-drop*.

pronominal acaba sendo traduzido. Em trabalhos futuros, seria interessante comparar o percentual de pronomes nominativos de O Pequeno Príncipe com o mesmo percentual em uma obra original do português da mesma época. Segundo Othero & Lazzari (2024), desde que o português brasileiro passou a simplificar as flexões verbais, utilizando a conjugação da terceira pessoa para a segunda pessoa (*você/vocês*), a elipse do sujeito vem diminuindo. Contudo, como a conjugação da segunda pessoa é usada no córpus em questão, poderia ensejar a elipse, como ocorre no português europeu.

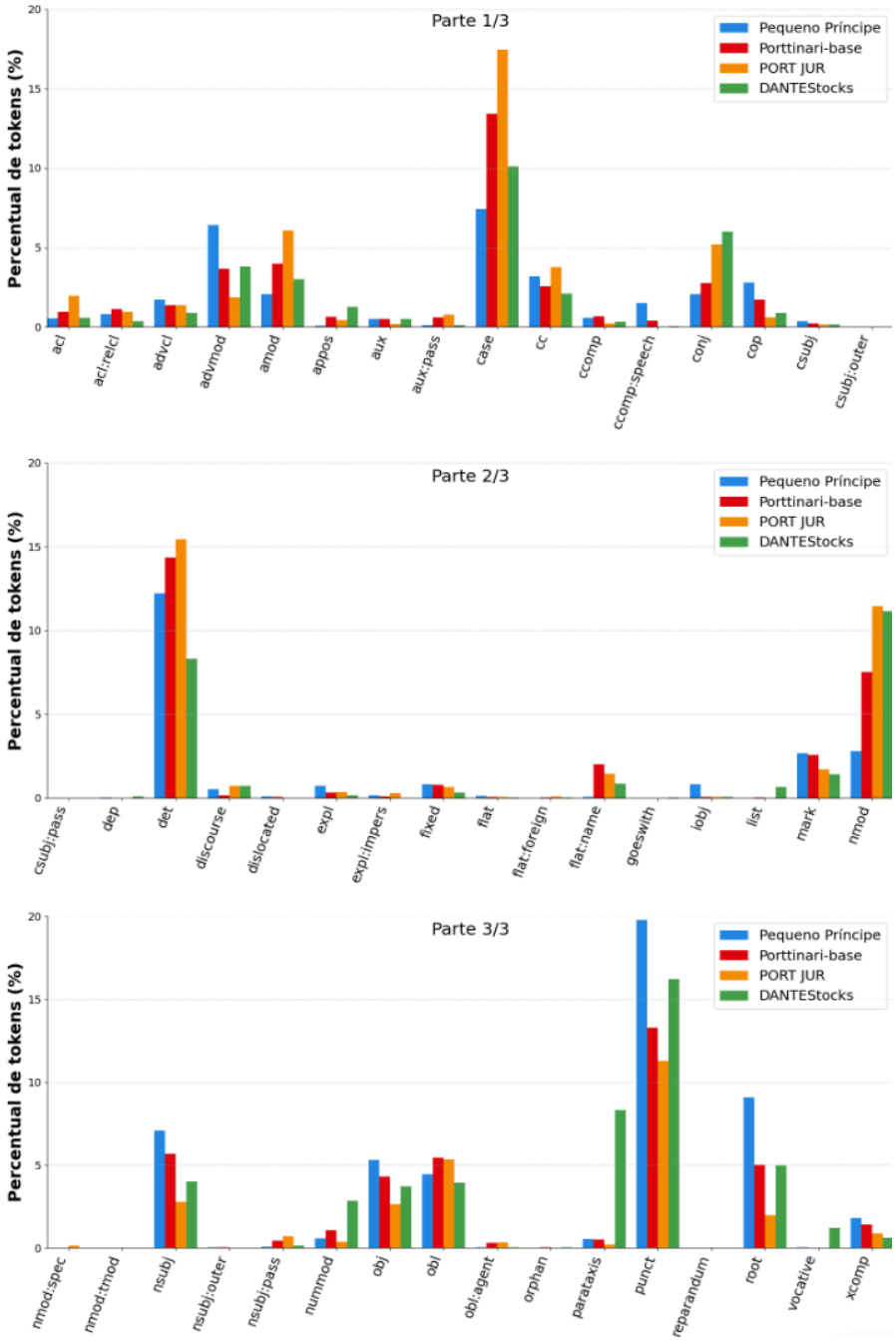


Figura 8. Distribuição das etiquetas DEPREL em diferentes córpus

Reunindo todos os dados, observamos que há correlação entre o fato de O Pequeno Príncipe ter a menor média de tamanho de sentença, e o fato de apresentar menor percentual de substantivos e modificadores nominais, menor percentual de preposições e maior percentual de verbos. Essas características conferem ao córpus maior simplicidade textual se comparado aos demais córpus de português na UD.

4. Considerações finais

A anotação sintática do córpus O Pequeno Príncipe é uma nova conquista dentro do objetivo de diversificar os gêneros disponíveis na coleção de córpus de português anotados com UD. São muitas as contribuições do gênero literário para a diversidade de textos. Uma delas é trazer muitas orações interrogativas, raras nos gêneros já anotados. Outra, é o uso expressivo da conjugação verbal de primeira e de segunda pessoas, no singular e no plural, característico do discurso direto. Além disso, duas marcas de diacronia são observadas no Pequeno Príncipe: o uso de "*tu*" e "*vós*" e o uso recorrente de pronomes clíticos, hoje pouco frequentes até no registro culto do português brasileiro.

Além de contribuir para aumentar o acervo do português em UD, o córpus O Pequeno Príncipe, por já ter sido anotado semanticamente por Anchiêta & Pardo (2018), abre a oportunidade para a exploração automática da correlação entre anotação sintática e anotação semântica. Para o leitor interessado, o córpus anotado e outras informações relevantes podem ser encontrados no portal web do projeto POeTiSA, em https://sites.google.com/icmc.usp.br/poetisa/resources-and-tools.

Agradecimentos

Este trabalho foi realizado no âmbito do Centro de Inteligência Artificial da Universidade de São Paulo (C4AI - http://c4ai.inova.usp.br/), com o apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM. Este projeto também foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei N. 8.248, de 23 de outubro de 1991, no âmbito do PPI-Softex, coordenado pela Softex e publicado como Residência em TIC 13, DOU 01245.010222/2022-44. Também se agradece ao apoio do INCT TILD-IAR (Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação - https://tildiar.dcc.ufmg.br/) (processo CNPq/SECTICS/CAPES/FAPs #408490/2024-1).

Referências

Anchiêta, R. and Pardo, T. (2018). Towards AMR-BR: A SemBank for Brazilian Portuguese language. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (pp. 449–460). European Language Resources Association (ELRA).

Andrade, P. L. C. de, Silva, R. M. and Pardo, T. A. S. (2026). Caracterização lexical e sintática de notícias falsas em português produzidas por humanos e por máquinas. In Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 2 (pp. 148–158). Salvador, Brazil.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M. and Schneider, N. (2013). Abstract meaning representation for sembanking. In Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse (pp. 178–186). Sofia, Bulgaria.

Di Felippo, A. and Roman, N. T. (2025). DANTEStocks: A multi-layered annotated corpus of stock market tweets for Brazilian Portuguese. Revista Brasileira de Linguística Aplicada, 25(1), 1–32.

Duran, M. S., Nunes, M. G. V., Lopes, L. and Pardo, T. A. S. (2022). Manual de anotação como recurso de processamento de linguagem natural: O modelo Universal Dependencies em língua portuguesa. Domínios de Lingu@gem, 16(4).

Duran, M., Lopes, L., Nunes, M. G. and Pardo, T. (2023). The dawn of the Porttinari multigenre treebank: Introducing its journalistic portion. In Proceedings of the 14th Brazilian Symposium in Information and Human Language Technology (pp. 124–133). Belo Horizonte, Brazil.

Duran, M. S., de Souza, E. A., Nunes, M. G. V., Pagano, A. S. and Pardo, T. A. S. (2025). Extending the enhanced Universal Dependencies: Addressing subjects in pro-drop languages. In Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025) (pp. 143–152). Ljubljana, Slovenia.

Fontes, M. G. (2012). A clivagem do constituinte interrogativo em sentenças interrogativas do português brasileiro: uma abordagem diacrônica. Signum: Estudos da Linguagem, 15(3), 149–170.

Freitas, C. and Pardo, T. A. S. (2025). PropBanks e representações semânticas: O que temos, o que queremos e o que podemos. Linguamática, 17(2), 3–31.

Guibon, G., Courtin, M., Gerdes, K. and Guillaume, B. (2020). When collaborative treebank curation meets graph grammars. In Proceedings of the Twelfth Language Resources and Evaluation Conference (pp. 5291–5300). Marseille, France.

Leal, S. E.; Duran, M. S.; Scarton, C. E.; Hartmann, N.; Aluísio, S. M. (2023) NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese. Lang Resources & Evaluation

Lopes, L., Duran, M. S. and Pardo, T. A. S. (2023). Verifica-UD: A verifier for Universal Dependencies annotation for Portuguese. In Proceedings of the 2nd Edition of the Universal Dependencies Brazilian Festival (pp. 451–460). Belo Horizonte, Brazil.

Lopes, L. and Pardo, T. (2024). Towards Portparser: A highly accurate parsing system for Brazilian Portuguese following the Universal Dependencies framework. In Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1 (pp. 401–410). Santiago de Compostela, Galicia/Spain.

Lopes, L., Nunes, M. G. V., Duran, M. S. and Pardo, T. A. S. (2025). A sintaxe no tribunal: Apresentando e explorando um corpus jurídico em português anotado sintaticamente segundo o modelo Universal Dependencies. In Proceedings of the XVI Symposium in Information and Human Language Technology (STIL) (pp. 220–232). Fortaleza/CE, Brazil.

de Marneffe, M.-C., Manning, C. D., Nivre, J. and Zeman, D. (2021). Universal Dependencies. Computational Linguistics, 47(2), 255–308.

Morais, M. A. T.; Salles, H. M. M. (2024). Resistência do dativo de primeira pessoa na batalha (quase) perdida dos clíticos pronominais do português brasileiro. Revista de Estudos da Linguagem, [S. l.], v. 30, n. 4, p. 1621–1657, 2024.

Oliveira, L., Claro, D. B. and Souza, M. (2023). DptOIE: A Portuguese open information extraction based on dependency analysis. Artificial Intelligence Review, 56, 7015–7046.

Othero, G. A.; Lazzari, M. (2024) Sujeitos e objetos nulos em português brasileiro: correlações e mudança. Revista de Estudos da Linguagem, [S. l.], v. 30, n. 4, p. 1831–1854, 2024.

Pagano, A. S., Rassi, A. and Pagano, A. C. S. (2026). A ordem e a função das palavras em uma sentença: Sintaxe. In H. M. Caseli and M. G. V. Nunes (Eds.), Processamento de linguagem natural: Conceitos, técnicas e aplicações em português (Vol. 1, 4th ed., chap. 6). BPLN.

Pardo, T., Duran, M., Lopes, L., Di Felippo, A., Roman, N. and Nunes, M. G. (2021). Porttinari: A large multi-genre treebank for Brazilian Portuguese. In Proceedings of the 13th Brazilian Symposium in Information and Human Language Technology (pp. 1–10). Porto Alegre, Brazil.

Qi, P.; Zhang, Y.; Zhang, Y.; Bolton, J.; Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In Association for Computational Linguistics (ACL) System Demonstrations. 2020. https://nlp.stanford.edu/pubs/qi2020stanza.pdf

Rademaker, A., Chalub, F., Real, L., Freitas, C., Bick, E. and de Paiva, V. (2017). Universal Dependencies for Portuguese. In Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017) (pp. 197–206). Pisa, Italy: Linköping University Electronic Press.

Souza, E., Silveira, A., Cavalcanti, T., Castro, M. and Freitas, C. (2021). PetroGold: Corpus padrão ouro para o domínio do petróleo. In Proceedings of the 13th Brazilian Symposium in Information and Human Language Technology (pp. 29–38). Porto Alegre, Brazil.

Straka, M. (2018). UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In: Proceedings of CoNLL 2018: The SIGNLL Conference on Computational Natural Language Learning, pp. 197-207, Association for Computational Linguistics, Stroudsburg, PA, USA.

Classificação automática de sentenças faladas com arquitetura híbrida: análise de estruturas topicalizadas, anti-topicalizadas e canônicas

Gerson Vitor P. Fernandes[1], Vitoria Maria A. Silva[2], Cid Ivan C. Carvalho[3]

[1] Universidade do Estado do Rio Grande do Norte (UERN) - Mossoró, RN - Brasil
[2] Universidade Federal Rural do Semi-Árido (UFERSA) - Mossoró, RN - Brasil
[3] Universidade Federal Rural do Semi-Árido (UFERSA) - Mossoró, RN - Brasil

prof.gersonvitor@gmail.com , cidivanc@ufersa.edu.br, vitoriamasmas@gmail.com

Abstract. *This article investigates the performance of an automatic classification system for spoken Brazilian Portuguese sentences considering canonical, topicalized, and antitopicalized structures using a hybrid architecture based on Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs). The research integrates syntactic, prosodic, and computational contributions to contextualize the corpus and define the scope of sentence processing; However, the study focuses on the experimental evaluation of a hybrid neural network architecture applied to the automatic classification of these sentences. The analyzed corpus consists of 612 declarative sentences recorded in a controlled environment, produced by 20 Brazilian Portuguese speakers. The methodology is quantitative and experimental, employing supervised learning and evaluation based on metrics such as precision, recall, F-score, and accuracy, using a confusion matrix. Tests were conducted with various hyperparameter configurations, varying convolutional filters, LSTM units, dropout rates, and the number of training epochs. The results demonstrated that more robust models exhibited a greater capacity to memorize training data but also a higher tendency toward overfitting. Conversely, leaner architectures combined with higher regularization rates achieved better generalization capabilities and greater stability during validation. The best performance was observed in the scenario with 30% dropout, which showed a significant reduction in the generalization gap between training and testing. It is concluded that the combination of RNNs and CNNs shows promise for the automatic classification of spoken sentences, although limitations regarding model generalization to unseen data still exists.*

Resumo. *O presente artigo investiga o desempenho de um sistema de classificação automática de sentenças faladas do português brasileiro, considerando estruturas canônicas, topicalizadas e antitopicalizadas, por meio de uma arquitetura híbrida baseada em Redes Neurais Recorrentes (RNNs) e Redes Neurais Convolucionais (CNNs). A pesquisa articula contribuições sintáticas, prosódicas e computacionais com a finalidade de contextualizar o corpus e delimitar o tratamento das sentenças faladas, porém o estudo concentra-se na avaliação experimental de uma arquitetura híbrida de redes neurais aplicada à classificação automática dessas sentenças. O corpus analisado é composto por 612 sentenças declarativas gravadas em ambiente controlado, produzidas por 20 falantes do português brasileiro. A metodologia caracteriza-se como quantitativa e experimental, utilizando aprendizagem supervisionada e avaliação baseada em métricas como precisão, recall, f-score e acurácia, por meio da matriz de confusão. Foram realizados testes com diferentes configurações de hiperparâmetros, variando filtros convolucionais, unidades LSTM, taxas de Dropout e número de épocas de treinamento. Os resultados demonstraram que modelos mais robustos apresentaram maior capacidade de memorização dos dados de treino, mas também maior tendência ao overfitting. Em contrapartida, arquiteturas mais enxutas associadas a maiores taxas de regularização obtiveram melhor capacidade de generalização e maior estabilidade durante a validação. O melhor desempenho foi*

observado no cenário com 30% de Dropout, em que houve redução significativa do hiato de generalização entre treino e teste. Conclui-se que a combinação entre RNNs e CNNs mostra-se promissora para a classificação automática de sentenças faladas, embora ainda existam limitações relacionadas à generalização do modelo em dados não vistos.

Palavras-chave: Modelo de classificação. Redes neurais. Arquitetura Híbrida. Frases faladas.

1. Introdução

Nas últimas décadas, os avanços no campo da Linguística Computacional e da Inteligência Artificial têm possibilitado o desenvolvimento de sistemas capazes de processar, reconhecer e classificar dados linguísticos de forma automatizada. Entre essas possibilidades, destaca-se o processamento automático da fala, área que reúne contribuições da linguística e da computação para a investigação de padrões linguísticos por meio de modelos computacionais treinados para reconhecimento e classificação de estruturas da língua.

Em estudos que envolvem dados de fala, aspectos fonético-fonológicos e prosódicos também podem ser incorporados à análise, especialmente quando se investigam relações entre entoação, organização sintática e processamento acústico. No âmbito da fala, a prosódia exerce papel fundamental na organização e interpretação dos enunciados, uma vez que elementos como entoação, duração, intensidade e ritmo contribuem para a construção de sentidos e para a distinção entre diferentes configurações sintáticas.

No português brasileiro, além das estruturas consideradas canônicas pela tradição gramatical, observa-se ampla produtividade de construções topicalizadas e antitopicalizadas, especialmente na modalidade oral, evidenciando a flexibilidade organizacional da língua e a relação entre sintaxe e prosódia. Os estudos sobre topicalização no português brasileiro ganharam destaque a partir das discussões de Pontes (1987), que apontam a coexistência de estruturas orientadas tanto para o sujeito quanto para o tópico. Posteriormente, pesquisas como as de Berlinck, Duarte e Oliveira (2009) ampliaram essa discussão ao sistematizar diferentes tipos de construções de tópico marcado. Paralelamente, investigações recentes têm buscado compreender de que maneira tais estruturas podem ser identificadas automaticamente por sistemas computacionais, sobretudo a partir de dados de fala.

Nesse contexto, o uso de redes neurais artificiais tem se mostrado promissor em tarefas relacionadas ao reconhecimento de padrões linguísticos e acústicos. Modelos baseados em aprendizado de máquina vêm sendo empregados em diferentes aplicações envolvendo a fala, como transcrição automática, detecção de distúrbios vocais e classificação de sentenças faladas. Entre essas arquiteturas, as Redes Neurais Recorrentes (RNNs) destacam-se pela capacidade de lidar com informações sequenciais e temporais, característica particularmente relevante para o processamento de sinais de fala.

Diante desse cenário, o presente estudo busca responder à seguinte questão: considerando as estruturas sintáticas canônicas, topicalizadas e anti-topicalizadas, quais apresentam melhor e pior desempenho em um sistema de classificação automática de sentenças faladas baseado em redes neurais recorrentes? Diante dessa questão, tem-se o seguinte objetivo geral: analisar o desempenho da classificação automática de frases faladas por meio de uma abordagem híbrida de redes neurais, considerando as categorias topicalizadas, anti-topicalizadas e canônicas.

Para isso, utiliza-se um corpus composto por sentenças declarativas do português brasileiro previamente gravadas em ambiente controlado e organizadas para

treinamento e validação do modelo computacional. A pesquisa articula três interfaces: sintaxe, prosódia e computação, buscando contribuir tanto para os estudos linguísticos acerca da organização sentencial do português brasileiro quanto para o desenvolvimento de modelos automáticos de classificação da fala.

Após esta introdução, apresentam-se os pressupostos teóricos que fundamentam a pesquisa, contemplando discussões acerca das construções topicalizadas, da prosódia e das redes neurais recorrentes e redes neurais convolucionais. Em seguida, descrevem-se os procedimentos metodológicos adotados para a constituição do corpus, treinamento do modelo e análise dos resultados.

2. Fundamentação teórica

2.1 Sentenças topicalizadas, anti-topicalizadas e canônicas

Em uma perspectiva normativa do português brasileiro, é consensual a adoção da ordem *sujeito + verbo + objetos + complementos* como o padrão organizacional mais comum dos elementos sentenciais. No entanto, na modalidade oral, outros padrões organizacionais são igualmente produtivos, como é o caso das sentenças de tópico. De acordo com Kenedy (2014), essas estruturas ocorrem quando um determinado constituinte (o tópico) é posicionado à periferia esquerda de uma sentença e, após essa colocação, há um comentário sobre esse elemento, como é o caso do seguinte exemplo: "*o vestido vermelho, a moça de preto comprou",* em que o objeto direto da sentença é posto na posição de tópico e o restante da sentença é organizado posteriormente. Desse modo, compreende-se que os constituintes podem se movimentar na sentença, criando outras estruturas, com a mesma correspondência semântica, no entanto, nem sempre com a mesma ênfase e/ou foco prosódico (Carvalho, Oliveira e Silva, 2025).

A discussão sobre a produtividade dessas construções teve início a partir de Li e Thompson (1976), que classificaram as línguas naturais em quatro tipologias básicas, a saber: i) línguas orientadas para a sentença - com proeminência de sujeito; ii) línguas orientadas para o discurso - com proeminência de tópico; iii) línguas com proeminência de tópico e sujeito; e iv) línguas que não têm proeminência de tópico nem de sujeito (sincréticas). No Brasil, as estruturas topicalizadas ganharam notoriedade com os estudos apresentados por Pontes (1987), que classificaram o português como uma língua de tópicos e sujeitos, haja vista a produtividade das sentenças topicalizadas.

Segundo Berlinck, Duarte e Oliveira (2009), as construções de tópico marcado não formam um grupo homogêneo. Com isso, é possível classificar o tópico marcado em, pelo menos, cinco tipos: 1) tópico pendente; 2) deslocamento à esquerda; 3) topicalização; 4) tópico-sujeito e 5) antitópico. No primeiro tipo, há conectividade semântica entre os constituintes e a sentença-comentário, mas não há conectividade sintática, como em "*drama, já basta a vida"*. No segundo tipo de tópico, um constituinte na posição de tópico tem um correferente expresso na sentença comentário, como em "cada elemento, *ele* possui o seu conjunto". O terceiro tipo envolve o movimento de um constituinte de sua posição de origem para a posição de tópico, deixando a posição original "vazia", como em "*carne, aqui em casa nós fazemos de várias formas",* em que o objeto direto é deslocado para a posição de tópico. O quarto tipo é aquele em que o tópico se encontra numa sentença com um sujeito nulo não-argumental, como em "meu celular acabou a bateria", em que parte de um constituinte é deslocado para a posição topicalizada. O último tipo de tópico apresentado é o antitópico que aparece à direita da sentença e ocupa uma posição

não-argumental e externa, como em “leva azeite de dendê, o acarajé”.

Neste artigo, o modelo de classificação considerou três possibilidades de organização, a saber: a organização considerada canônica pela tradição gramatical e pelos estudos linguísticos (sujeito + verbo + objeto); a segunda é a topicalização e a antitopicalização, na perspectiva de Berlinck, Duarte e Oliveira (2009).

Nesta seção, foi apresentado o conceito de tópico e suas diferentes classificações. Na seção seguinte, serão apresentados os conceitos relacionados à prosódia.

2.2 Prosódia

O termo “prosódia” deriva do grego προσῳδία, no qual o prefixo προσ significa "sobre" e a raiz ᾠδή, "canção". Originalmente, conforme explicam Rebollo-Couto e Seara (2019), o termo referia-se à "canção sobre os sons segmentais", sendo utilizado para descrever os acentos das palavras. Com o tempo, seu significado foi sendo expandido e passou a abranger uma gama mais ampla de elementos ligados à fala humana.

Assim, a prosódia passa a compreender um conjunto de fenômenos que envolvem parâmetros como altura, intensidade, duração, pausas e velocidade da fala, o estudo da entoação, acento e ritmo nas línguas naturais (Scarpa, 1999). De maneira geral, entende-se que a prosódia não está relacionada diretamente ao conteúdo da fala, mas ao modo “como” a fala é produzida, refletindo a maneira de transmitir significados por meio da variação sonora (Barbosa, 2019).

No estudo de Carvalho, Oliveira e Silva (2025), foi utilizada uma perspectiva prosódico-sintática, em que se observou a estrutura sintática e a sua realização na fala. Nesse estudo, elementos como a frequência fundamental (F0) e o contorno entoacional foram analisados, a fim de identificar diferenças na região nuclear do enunciado; o estudo apontou diferenças no contorno entre falantes masculinos e femininos.

O presente estudo articula-se entre três interfaces: 1. sintaxe, 2. prosódia e 3. computacional. A sintática relaciona-se pelas construções trabalhadas na seção 2.1: canônicas, topicalizadas e anti topicalizadas. O nível prosódico justifica-se pelo fato de trabalharmos com sentenças faladas. Embora a proposta deste estudo não seja realizar uma análise acústica dessas sentenças, elas foram gravadas em local controlado, devidamente tratadas e segmentadas, a fim de garantir a melhor qualidade para a análise computacional. Por fim, a interface computacional justifica-se por um sistema de classificação, adotando um modelo híbrido, conforme discutido na seção 2.3.

2.3 Classificação automática e redes neurais recorrentes

No campo da linguística computacional, o processamento automático da fala tem sido investigado em diálogo com a engenharia computacional, principalmente nos estudos envolvendo *o aprendizado de máquina* e *processamento de sinais*. Com base no estudo de Almeida *et al.* (2024), nota-se que os modelos treinados podem ser usados como uma ferramenta de transcrição da fala de indivíduos com gagueira. De forma semelhante, Azevedo (2025) em sua análise na identificação de distúrbios da voz, entre eles a *disfonia, laringite, pólipo e paralisia*, o autor apresentou dados positivos, no teste de validação, o modelo atingiu uma taxa de 90%, mesmo com algumas limitações.

Entretanto, através de estudos computacionais é possível realizar uma abordagem sintática, linguística e computacional, como no caso de Carvalho, Oliveira e Silva (2025), quel, baseados no modelo de redes neurais convolucionais, estudaram a classificação automática de sentenças faladas, através de espectrogramas. Mesmo alcançando uma acurácia abaixo dos 90%, o modelo apresentou uma excelente precisão em alguns tipos de sentenças.

O aprendizado de máquina parte do princípio que as máquinas conseguem aprender como uma "mente" humana, mesmo produzida artificialmente. É importante ressaltar que ainda existem muitas limitações, porém os modelos computacionais conseguem atingir excelentes resultados. Dijkstra (2021) cita três tipos de aprendizagem em seu estudo: o *supervisionado*, *não supervisionado e o reforço*. Cada treinamento tem uma abordagem específica, portanto, é necessário entender para qual finalidade o modelo será usado, a fim de encontrar o treino ideal.

No primeiro caso mencionado, o sistema aprende através de um banco de dados previamente "alimentado". A máquina tenta encontrar padrões, e, quanto mais dados tiver para trabalhar, mais fácil o reconhecimento. O segundo caso tenta encontrar e criar grupos de dados; mesmo não existindo um direcionamento inicial, ele analisa as características de cada uma, identifica padrões e agrupa no local que mais se aproxima com o objeto. O terceiro caso é baseado em um sistema de erros e acertos, recompensas e punições em um determinado ambiente; através de falso positivo e falso negativo, trata-se de um modelo que se ajusta de acordo com um direcionamento.

Assim como os tipos de aprendizado, percebe-se a necessidade de distinguir a diferença entre redes neurais recorrentes e convolucionais, pois cada rede tem uma abordagem diferente. Faceli (2011) explica que as redes neurais aproximam-se de uma tentativa artificial de simular como o sistema neural funciona biologicamente, sendo uma tentativa de automatizar tarefas cotidianas por meio de máquinas treinadas, que incluem neurônios artificiais. O conjunto deles forma as redes neurais artificiais (RNAs).

As redes neurais funcionam com arquiteturas diferentes, cada uma delas podendo ser mais adequada para tarefas específicas, como no caso das RNNs que trabalham com sequência e temporalidade, diferente das CNNs que conseguem encontrar padrões em múltiplas categorias como imagens, representações acústicas, espectrogramas, matrizes e outros.

No presente estudo, utilizamos uma abordagem híbrida usando as Redes Neurais Recorrentes (RNNs) e Redes Neurais Convolucionais para o sistema de classificação de sentenças faladas. Em primeiro caso, o estudo de Moura (2018), afirma que as redes neurais recorrentes são formadas por uma arquitetura de redes em camadas, que funciona da seguinte forma: inicialmente o sistema captura as relações temporais por meio do compartilhamento de parâmetros e a arquitetura considera as informações anteriores, em diferentes momentos do processamento sequencial, auxiliando na predição.

Por outro lado, as convolucionais distinguem-se por conseguir identificar padrões em dados estruturados em matrizes, imagens, espectrogramas ou até mesmo representações acústicas da fala. Contudo, sentenças faladas não são analisadas de forma bruta: é necessário transformá-las em alguma das variáveis mencionadas anteriormente, por meio de representações visuais ou numéricas. Conforme Dijkstra (2021), esse tipo de arquitetura é formada por diversas camadas, entre elas as de convolução e pooling.

É importante ressaltar que a forma de aprendizado que mais se aproxima com a nossa pesquisa é a aprendizagem supervisionada, pois o nosso objetivo é a identificação de três categorias sintáticas com características específicas, por meio do treinamento do sistema, espera-se resultados satisfatórios. Por fim, utilizamos uma abordagem híbrida, com um modelo de classificação baseado no sistema de redes neurais recorrentes (RNNs) e de redes neurais convolucionais (CNNs).

3. Metodologia

Como mencionado anteriormente, o estudo não concentra-se em pistas prosódicas, como duração, pausa, sílaba tônica e ou elementos acústicos, porém, por se tratar de sentenças faladas, adotamos todos os princípios de coleta e qualidade usados em análises prosódicas, incluindo o rigor exigido para a escolha dos falantes, instrumentos, comitê de ética (CAAE 79118423.5.0000.9547) e ambiente controlado.

O corpus utilizado é o mesmo conjunto adotado por Carvalho, Oliveira e Silva (2025), com um total de 612 sentenças declarativas, arquivos gerados no formato .wav. No total, foram 20 participantes, entre eles 10 do gênero masculino e 10 do feminino, ambos com faixa etária entre 18 a 40 anos. Foram selecionados entrevistados que não tivessem dificuldades de fala, causadas por questões cognitivas ou por uso de aparelhos ortodônticos, critérios adotados por questões metodológicas.

Como exemplo das sentenças inseridas no modelo, tem-se, na estrutura canônica, "O diretor concordou com a reunião", topicalizada "Com a reunião, o diretor concordou" e antitopicalizada "Concordou com a reunião, o diretor".

A pesquisa enquadra-se como um estudo quantitativo experimental, por utilizarmos uma boa quantidade de sentenças, como também um sistema pelo Google Colab com arquitetura híbrida para discutir a quantidade de erros e acertos do sistema.

A métrica para validação do sistema foi a matriz de confusão, uma tabela que consegue avaliar o desempenho de um sistema, por meio das variáveis: *precisão, recall, f-score e acurácia.* Com esses parâmetros é possível definir a qualidade da arquitetura, entretanto, para encontrar os valores dos elementos mencionados anteriormente, trabalha-se por meio de quatro possibilidades. Segundo Monard e Baranauskas (2003, p. 48) as possibilidades são: "*Verdadeiros positivos (Tp), Falsos positivos (Fp), Falsos negativos (Fn) e Verdadeiros negativos (Tn)".* O primeiro caso ocorre quando o sistema entrega o resultado corretamente; o segundo entrega um resultado, porém falso, não condizente com a verdade; o terceiro, quando é esperado um resultado verdadeiro e ele entrega um negativo; o quarto ocorre quando é esperado uma informação negativa e ele realmente entrega.

4. Descrição e análise dos resultados

Neste estudo, foi adotado o termo época referente a cada ciclo completo de treinamento do modelo usando os dados disponíveis. Os hiperparâmetros são configurações feitas antes do treinamento, como filtros, unidades LSTM, taxa de Dropout e número de épocas. Inicialmente, a avaliação do modelo híbrido baseou-se na variação controlada dos hiperparâmetros de convolução, memória e regularização. Foram definidos dois cenários de testes principais (A e B). Para cada um deles, aplicou-se uma variação na taxa de Dropout dos valores de 20%, 25% e 30%. A estrutura detalhada de cada cenário encontra-se resumida na **tabela 1**.

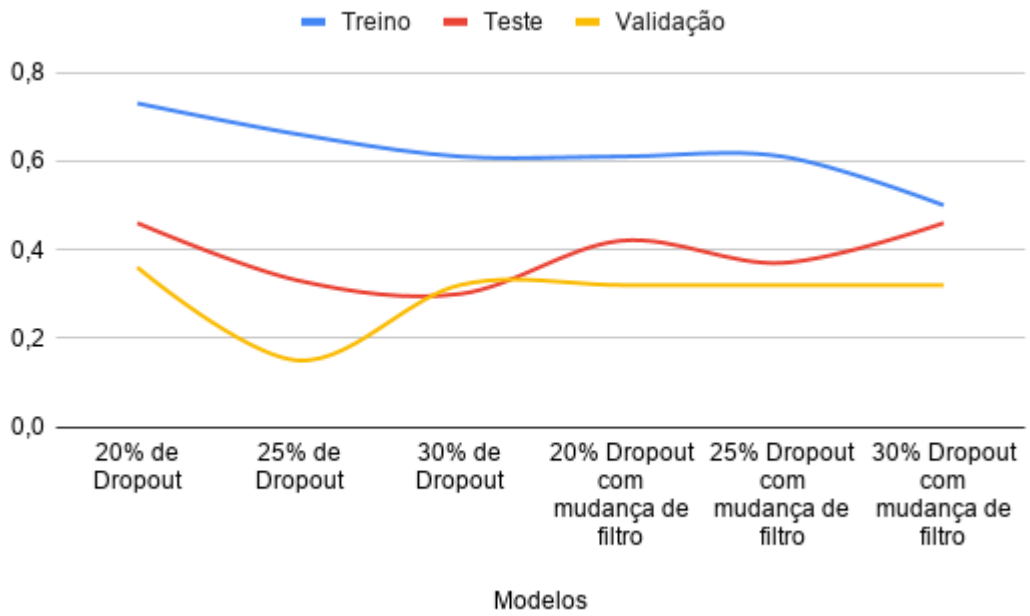
Tabela 01. **Apresentação dos cenários A e B considerando a mudança nos parâmetros da rede neural híbrida**

Componente do modelo	Parâmetros do cenário A	Parâmetros do cenário B
Conv1D (1ª Camada)	128	64
Conv1D (2ª Camada)	256	128
LSTM	128	64
Densa (Dense)	64	64
Dropout (Testados)	20%, 25%, 30%	20%, 25%, 30%

Fonte: Dados da pesquisa

Com base na mudança desses parâmetros, apresenta-se abaixo o **gráfico 1** com os resultados e a análise.

Gráfico 01. **Resultados obtidos do treino, do teste e da validação nos modelos de rede neural híbrida**



Fonte: Dados da pesquisa

Para os modelos do Cenário A, utilizou-se uma estrutura de parâmetros mais robustos com 128 e 256 filtros convolucionais e 128 de unidades LSTM. Essas redes demonstraram maior capacidade de memorização dos dados de treino, atingindo a acurácia máxima observada de aproximadamente 73% com 20% de Dropout. No entanto, o aumento do Dropout para 25% e 30% reduziu progressivamente o desempenho de treino e teste. Destaca-se uma instabilidade na validação quando a taxa de 25% de Dropout foi aplicada e a acurácia sofreu uma queda abrupta para cerca de 15%, sugerindo que o excesso de parâmetros associado a uma taxa intermediária de regularização prejudicou a convergência local do modelo.

No cenário B, em que se fez a redução da complexidade estrutural com 64 e 128 filtros e com 64 unidades LSTM, observou-se uma mudança significativa na dinâmica de aprendizado. O comportamento da curva de validação (linha amarela) tornou-se notavelmente estável, fixando-se de forma linear na faixa dos 32% independentemente da taxa de Dropout aplicada. Isso evidencia que a redução da capacidade da rede atenuou a variância extrema observada no cenário anterior.

Depois da escolha do modelo mais adequado ao trabalho com base nas informações acima, avaliou-se a dinâmica do modelo híbrido com base na convergência e na estabilidade a longo prazo, realizando um ensaio variando o número de épocas de treino de 50 até 200. Os resultados quantitativos de acurácia para os conjuntos de treino, teste e validação encontram-se sistematizados na **tabela 2**, e a sua correspondente evolução temporal está ilustrada no **gráfico 2**.

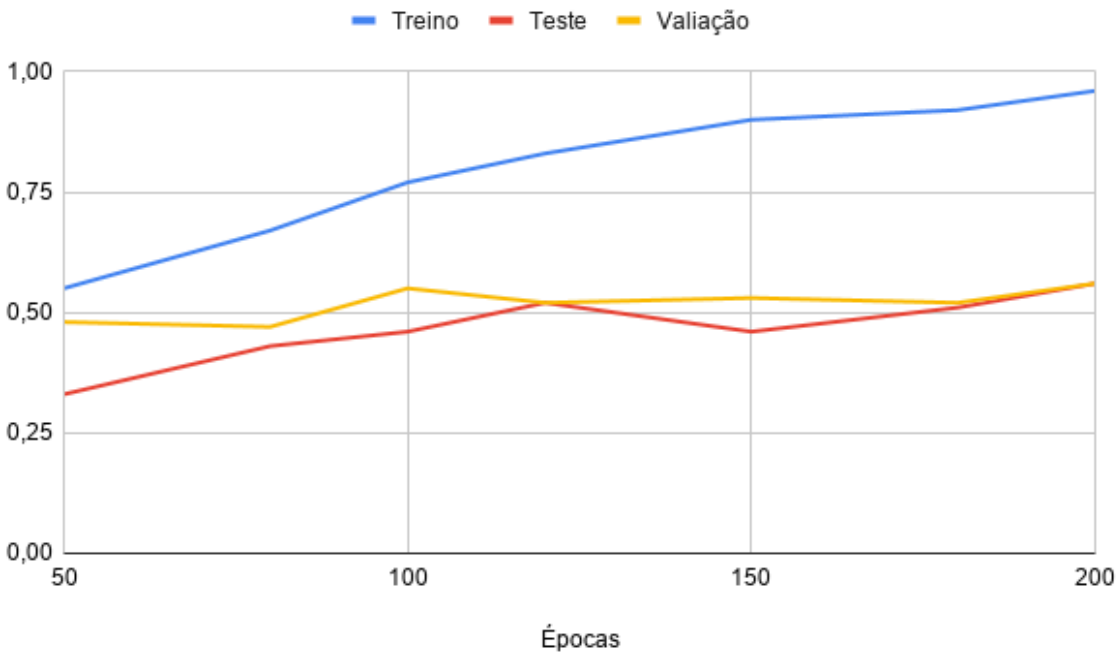
Tabela 02. **Análise da acurácia no comportamento do modelo com 30% de Dropout e com as mudanças nos valores dos filtros nas épocas de treinamento**

Épocas	Acurácia		
	Treino	Teste	Validação
50	0,55	0,33	0,48
80	0,67	0,43	0,47
100	0,77	0,46	0,55
120	0,83	0,52	0,52
150	0,90	0,46	0,53
180	0,92	0,51	0,52
200	0,96	0,56	0,56

Fonte: Dados da pesquisa.

A análise do gráfico mostra que há um crescimento acentuado da acurácia do treino, que inicia em 55% (50 épocas) e atinge o limiar de 96% ao fim de 200 épocas. Este comportamento evidencia a elevada capacidade adaptativa da arquitetura em mapear e memorizar os padrões intrínsecos dos dados de treino. No entanto, ao observar comparativamente os conjuntos de teste e de validação, constata-se que existe uma estagnação no desempenho a partir da época 100. Veja-se que, entre as épocas 120 e 200, a acurácia de validação se estabilizou num patamar entre 52% e 56%. Como consequência disso, o gráfico mostra um hiato de generalização (*generalization gap*), de forma que a diferença entre a acurácia de treino e as métricas de validação e de teste se expandiu, passando de 7% às 50 épocas para 40% às 200 épocas.

Gráfico 02. **Ilustra os valores no comportamento do modelo com 30% de Dropout e com as mudanças no valores dos filtros nas épocas de treinamento**



Fonte: Dados da pesquisa.

Outro ponto que se deve destacar é o fato de que este cenário se caracteriza como um quadro de *overfitting*. À medida que o treino avança acima de 100 épocas, a rede deixa de extrair as características generalizáveis dos áudios e passa a ajustar-se ao ruído específico do conjunto de treino. Embora o pico máximo de validação e teste ocorra às 200 épocas (56%), o custo computacional e o risco de perda na qualidade da classificação tornam os modelos gerados após a época 120 menos viáveis para aplicação prática em cenários reais.

5. Conclusão

Quando se observa o ponto de destaque do experimento que ocorreu no cenário B com 30% de Dropout, vê-se que a acurácia de treinamento reduziu para 50%, enquanto a acurácia de teste alcançou seu segundo maior pico histórico (46%). Consequentemente, o hiato de generalização (*generalization gap*) entre o treino e o teste foi reduzido para apenas 4%. Este comportamento comprova a eficácia da combinação de uma arquitetura mais enxuta com uma maior regularização, resultando no modelo mais robusto, menos propenso ao *overfitting* e com melhor potencial de generalização entre os modelos testados.

É necessário um olhar atento em relação aos dados, pois mesmo apresentando um potencial positivo, os dados de acurácia não são satisfatórios, mas a proposta híbrida apresentou um potencial, pois não apenas memorizou os dados do treino, mas conseguiu classificar, mesmo que ainda não possa ser usado com segurança. Portanto, os resultados do estudo apresentaram um caminho promissor na classificação de sentenças faladas, principalmente por utilizar parâmetros equilibrados no treino e teste do sistema, abrindo possibilidades para novos ajustes e abordagens e melhorias.

Referências

ALMEIDA, Rodrigo José S. et al. Aprendizado de máquina no apoio à transcrição e classificação da fala gaguejada: uma revisão sistemática da literatura. *Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)*, p. 400-411, 2024.

AZEVEDO, Gabriel Almeida et al. *Análise acústica da fala para auxílio à detecção e à classificação de distúrbios da voz*. 2025.

Barbosa, P. A. (2019). *Prosódia*, São Paulo, Parábola.

Barbosa, P. A. (2022). *Manual de Prosódia Experimental*, 1 ed, Editora da Abralin.

Berlinck, R. A.; Duarte, M. E. L. and Oliveira, M. (2009). Predicação. In Castilho, A.

T., Kato, M. A., Nascimento, M., *Gramática do português culto falado no Brasil*, Campinas, Editora da Unicamp.

CARVALHO, Cid Ivan C.; OLIVEIRA, Francisca Ticiany B. L.; SILVA, Vitória Maria A.. *Modelo de Classificação Automática de Frases Faladas com Abordagem em Redes Neurais Convolucionais*. *In*: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (STIL), 16. , 2025, Fortaleza/CE. **Anais** [...]. Porto Alegre: Sociedade Brasileira de Computação, 2025 . p. 519-525. DOI: https://doi.org/10.5753/stil.2025.37852.

Casanova et al. (2024). Recursos para o Processamento de Fala. In Caseli, H. M. and Nunes, M.G.V. (org.) *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, 2 ed, BPLN. Disponível em: https://brasileiraspln.com/livro-pln/2a-edicao.

Catarino, M. H (2025). Redes Neurais. Rio de Janeiro, Editora Freitas Bastos.

DESHPANDE, Kedar et al. A Time-Distributed CNN-LSTM with Attention Model for Speech Based Emotion Recognition. In: *Proceedings of the 2024 7th International Conference on Digital Medicine and Image Processing*. 2024. p. 67-71.

DIJKSTRA, Bauke Alfredo. *Reconhecimento de fonemas utilizando redes neurais convolucionais para transcrição fonética automática*. 2021. Dissertação de Mestrado. Universidade Tecnológica Federal do Paraná.

FACELI, Katti et al. *Inteligência artificial: uma abordagem de aprendizado de máquina.* 2011.

KENEDY, Eduardo. O status tipológico das construções de tópico no Português Brasileiro: uma abordagem experimental. *Revista da ABRALIN*, 2014.

LADD, D. Robert; COUTO, Tradução de Letícia Rebollo; SEARA, Izabel Christine. O que é prosódia. *Working Papers em Linguística*, v. 20, n. 1, p. 8-46, 2019.

LI, Charles N.; THOMPSON, Sandra A. Subject and topic: A new typology of language. In: *Subject and topic*. Routledge, 1976. p. 457-489.

Lira, Z. (2009). *A entoação modal em cinco falares do Nordeste brasileiro*, Tese (Doutorado) - UFPB, João Pessoa.

Lucente, L. (2022). Notação Entoacional. In Oliveira Júnior, M. *Prosódia, Prosódias*, Editora Contexto, pages 45-66.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto. Conceitos sobre aprendizado de máquina. *Sistemas inteligentes-Fundamentos e aplicações*, v. 1, n. 1, p. 32, 2003.

MOURA, Gustavo Bennemann de. *Redes neurais recorrentes para a classificação de estruturas retóricas.* 2018.

Pontes, E. (1987). *O Tópico no Português do Brasil.* Campinas, Editora Pontes.

Raso, T.; Teixeira, B.; Barbosa, P. (2020). *Modelling automatic detection of prosodic boundaries for Brazilian Portuguese spontaneous speech.* Journal of Speech Sciences, Campinas, SP, v. 9, n. 00, pages 105–128. DOI: 10.20396/joss.v9i00.14957. Disponível em:

https://econtents.bc.unicamp.br/inpec/index.php/joss/article/view/14957.

SCARPA, E. M (Org). *Estudos de prosódia*. Campinas: Unicamp, 1999.

Srivastava, N. et al. (2014). *Dropout: A Simple Way to Prevent Neural Networks from Overfitting, Journal of Machine Learning Research,* 15, pages 1929-1958. https://jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf.

Wagner, A. (2008). *Automatic labeling of prosody, ITRW on Experimental Linguistics*, ExLing 2008, 25-27 August 2008, Athens, Greece. Disponível em: https://www.isca-archive.org/exling_2008/wagner08_exling.pdf.

Wibawa, I.D.G.Y.A; Darmawan, I.D.M.B.A (2021). *Implementation of audio recognition using mel frequency cepstrum coefficient and dynamic time warping in wirama praharsini, Journal of Physics*: Conference Series, DOI: 10.1088/1742-6596/1722/1/012014.

Wightman, C.W.; Ostendorf, M. (1994). *Automatic labeling of prosodic patterns, IEEE Transactions on Speech and Audio Processing*, 2(4), pages 469–481.

A Comparative Analysis of Adjectives in Image Descriptions in Brazilian Portuguese and English

Maucha Andrade Gamonal[1], Adriana Pagano[2], Felix Engler[2], André V. Lopes Coneglian[2], Ely E. S. Mattos[3]

[1]Universidade Estadual de Campinas, [2]Universidade Federal de Minas Gerais, [3]Universidade Federal de Juiz de Fora
mgamonal@unicamp.br, apagano@ufmg.br, felixenglera@gmail.com, coneglian@ufmg.br, ely.matos@ufjf.edu.br

Abstract. *This paper investigates how adjectives construe visual experience in Brazilian Portuguese and English descriptions from the Framed Multi30K dataset. Grounded in Systemic Functional Grammar and Frame Semantics, the study annotates adjectives as experiential Epithets, interpersonal Epithets, or Classifiers, and examines the frames they evoke. The results show that experiential Epithets predominate, suggesting that adjectives mainly construe visually salient properties. Frame annotation shows that* `Color` *is the most frequent frame. By identifying salient adjectival meanings in human-produced descriptions, the study provides evidence of which visual attributes tend to be prioritized by human observers.*

Resumo. *Este artigo investiga o modo como os adjetivos representam a experiência visual em descrições, em português brasileiro e em inglês, do conjunto de dados Framed Multi30K. Com base na Gramática Sistêmico-Funcional e na Semântica de Frames, o estudo classifica os adjetivos em Epítetos experienciais, Epítetos interpessoais e Classificadores, além de examinar os frames por eles evocados. Os resultados mostram que os Epítetos experienciais são predominantes, o que indica que os adjetivos são empregados principalmente para representar propriedades visualmente salientes. A anotação semântica também revela que* `Color` *é o frame mais frequente. Ao identificar os significados adjetivais mais salientes, o estudo contribui para compreender quais atributos visuais tendem a receber maior destaque por parte dos observadores.*

1. Introduction

Image descriptions play a central role in accessibility resources, such as audiodescription, enabling blind and visually impaired people to access visual content. However, they remain frequently absent in online environments: according to Movimento Web Para Todos (2024), only 42.83% of 26.3 million Brazilian websites include image descriptions. This gap has motivated the development of automatic description systems, which rely on computer vision and natural language generation to identify visual content and transform it into language. However, accessible description requires more than the recognition of entities in an image: it also involves decisions about which attributes and meanings are relevant enough to be verbalized. Improving such systems therefore depends on a better understanding of how human-produced descriptions are linguistically structured and

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 487–498
October 19-22, 2026 ©2026 Association for Computational Linguistics

which kinds of meanings they tend to foreground. Adjectives are central to this issue because they help construe visually salient properties, including color, size, age, material, posture, physical appearance, and evaluation.

This study investigates the use of adjectives in image descriptions in Brazilian Portuguese and English, based on a subset of the Framed Multi30K dataset [Viridiano et al. 2024]. The analysis focuses on 300 sentences describing 30 images, a sample selected to enable detailed manual annotation and qualitative control. Adjectives are classified following Systemic Functional Grammar [Halliday and Matthiessen 2013] and analyzed in relation to the entities they modify. In addition, Frame Semantics [Fillmore 1982] is used to examine the semantic frames evoked by these adjectives, with support from the LOME FrameNet parser [Xia et al. 2021], whose outputs are critically assessed.

As an exploratory study, this paper does not aim at large-scale generalization, but seeks to identify tendencies in adjective use in human-produced image descriptions and to discuss how adjectival meanings contribute to the selection and construal of visually salient attributes in multimodal and accessibility-oriented contexts. As [Coneglian and Pagano 2025] have shown, the Framed Multi30K image description constitute a prime terrain for the investigation of how lexicon and grammar are mobilized to construe meaning in language.

2. Theoretical Background

2.1. Systemic-Functional Grammar

Systemic-Functional Grammar [Halliday and Matthiessen 2013] views language as organized around three metafunctions: ideational, interpersonal, and textual. The ideational metafunction is associated with the representation of experience through processes, participants, and circumstances; the interpersonal metafunction with interaction, stance, and evaluation; and the textual metafunction with the organization of meanings in discourse.

The metafunctional perspective operates across the rank scale, from the clause to the morpheme, including the intermediate ranks of group/phrase and word. At each rank, units are analyzed in terms of both class and function: class identifies the type of unit, while function specifies the role that the unit performs in the structure of a higher-ranking unit. At group level, nominal groups typically function within the clause as participants, but they also have their own internal organization. Within the nominal group, words and word groups perform different functions, such as identifying, classifying, describing, or quantifying the entity being referred to. Among the distinct words classes that make up a nominal group, adjectives may function as Epithets, contributing descriptive or evaluative meanings, or as Classifiers, helping to categorize the entity as a particular type [Halliday and Matthiessen 2013]. Epithets express qualities attributed to an entity and may be either experiential, describing objective properties, or interpersonal, expressing subjective evaluation. Classifiers, in contrast, define subclasses of entities and are not typically gradable. One of the sentences in our corpus illustrates these different functions:

(1) A skier is overlooking the beautiful white snow-covered landscape.

In this example, *beautiful* functions as an interpersonal Epithet, expressing subjective evaluation, while *white* and *snow-covered* function as experiential Epithets, des-

cribing objective properties of the entity.

2.2. Frame Semantics

Frame Semantics [Fillmore 1982] assumes that word meaning is understood in relation to structured background knowledge, or frames, which are evoked during language use. Each frame represents a conceptual scenario involving specific participants and relations, enabling a detailed account of meaning beyond general semantic roles. In this study, frame semantics is used to examine the semantic frames evoked by adjectives in image descriptions. Frame annotation is supported by the LOME FrameNet parser [Xia et al. 2021].

A single sentence may evoke multiple frames. For example:

(2) The old man bought a pie.

In this sentence, *bought* evokes a `Commerce` frame, *pie* evokes a `Food` frame, and the adjective *old* evokes an `Age` frame[1].

This illustrates how adjectives contribute to the activation of specific domains of meaning within a broader semantic structure.

Adjectives are therefore analyzed in terms of the frames they evoke—such as `Color`, `Age`, or `Movement`—highlighting how different domains of meaning are activated in each language.

By combining functional and frame-semantic perspectives, this study identifies cross-linguistic tendencies in adjective use and contributes to the linguistic description of Brazilian Portuguese in multimodal contexts, as well as to discussions on semantic annotation practices.

2.3. A comparative typology of adjective lexemes in Brazilian Portuguese and English

Adjective lexemes pose an interesting challenge for inter-language comparison. In comparing Brazilian Portuguese, there are three aspects of adjectives lexemes that must be considered: (i) the lexical nature of the class; (ii) the grammatical behavior of such lexemes in each language; (iii) other language-specific means through which 'adjectival functions' may be realized.

First, Brazilian Portuguese and English are languages that have an open class of adjectives, and these are languages that present a fairly robust morphological derivational paradigm for the coinage of adjective words from nouns and verbs. Secondly, adjective lexemes differ in their grammatical behavior in both languages: while Brazilian Portuguese adjectives are grammatically aligned with nouns, English adjectives present a behavior that differs from both nouns and verbs [Dixon 2010]. This difference between the two languages poses a significant challenge in comparing captions of image descriptions.

Thirdly, both languages present different constructions that express the adjectival function of modification of an entity. In both languages, participles may be lexicalized

[1]By convention, italics are used in this paper both for emphasis on lexical items discussed in the analysis and for English translations of sentences originally written in Brazilian Portuguese. Monospaced font is used to represent frame names exactly as they appear in the FrameNet database

into functioning as modifying adjectives. And, most significantly, both languages present a construction [Preposition + Noun] that functions as a modifier of a nouns, as in 'mesa **de madeira**', *table **of wood***, in Brazilian Portuguese. These cases will not be considered in our analysis, since our focus is on the lexical expression on modification.

3. Resources used

3.1. FrameNet Brasil

FrameNet Brasil [Torrent et al. 2022] is a lexicographic annotation project based on Berkeley FrameNet, organized around the description of meaning through frames and their associated elements. The resource includes lexical units, frames, frame elements, and relations between frames, as well as extensions to multimodal semantic annotation. In recent years, FrameNet Brasil has also contributed to the development of multimodal datasets and annotation methodologies integrating textual and visual information [Viridiano et al. 2024] [Belcavello et al. 2024] , expanding the application of Frame Semantics to image and audiovisual description contexts.

Lexical units correspond to pairings of a word with a specific meaning that evokes a frame. For example, the adjective *white* may evoke different frames depending on its meaning, such as `Color` or `Race descriptor`. Frames represent structured conceptual scenarios, including participants and relevant attributes. Each frame is associated with a set of frame elements, which encode semantic roles and are typically classified as core, peripheral, or extra-thematic [Ruppenhofer et al. 2016].

In this study, FrameNet Brasil provides the semantic framework used to analyze meanings evoked by adjectives in image descriptions, supporting the investigation of cross-linguistic semantic variation in multimodal contexts.

3.2. LOME FrameNet Parser

LOME FrameNet parser [Xia et al. 2021] is a multilingual information extraction system designed to identify semantic frames, frame elements, and their corresponding trigger spans directly from raw text. Unlike parsers conditioned on predefined targets, LOME performs full-structure prediction, extracting all frames and roles present in a sentence. The model was trained on annotated data in multiple languages, including Portuguese and English.

In this study, LOME was used to re-analyze the selected image descriptions in both languages, with the specific goal of identifying frames evoked by adjectives. Its outputs were subsequently reviewed to ensure analytical reliability, particularly in cases of misidentification or omission. This procedure enabled both the examination of frame distribution across languages and the assessment of the parser's performance in a multimodal annotation context.

3.3. Framed Multi30k dataset

Framed Multi30K [Viridiano et al. 2024] is a multimodal dataset that extends Flickr30K by incorporating multilingual descriptions and frame-semantic annotations. The original Flickr30K dataset contains 31,783 images of everyday scenes, each paired with five English descriptions. Framed Multi30K expands this resource by including both translated

and original descriptions in Brazilian Portuguese, as well as frame and frame element annotations aligned with visual regions.

An example from the dataset is presented in Figure 1 with two English and two Brazilian Portuguese descriptions of the same image:



Figura 1. Example of image descriptions (Flickr30K 1003420127)

(3) A group of adults, inside a home, sitting on chairs arranged in a circle, playing musical instruments.

(4) Five musicians, a man and four women, practicing sheet music (using flutes) in a living room.

(5) Cinco pessoas em uma sala tocam flauta sentadas em cadeiras.
(Five people in a room play the flute sitting on chairs.)

(6) Um grupo de músicos reunidos em círculo estão sentados em cadeiras em uma sala clara e residencial segurando instrumentos e próximos a estantes de partitura.
(A group of musicians gathered in a circle are sitting on chairs in a light, residential room holding instruments and near sheet music stands.)

4. Methodology

A sample of 30 images was randomly selected from the Framed Multi30K dataset [Viridiano et al. 2024]. The dataset includes five original descriptions in English and five in Brazilian Portuguese per image, resulting in a total of 300 sentences. The selection of this subset allows for detailed manual annotation and controlled comparison across languages, prioritizing analytical depth over large-scale generalization.

All descriptions were organized on a spreadsheet. Adjectives were manually identified and classified according to Systemic Functional Grammar [Halliday and Matthiessen 2013] into experiential Epithets, interpersonal Epithets, and Classifiers. In addition, adjectives were annotated according to the type of entity they modify (person, clothing, object, location, animal, or other).

Following this step, all sentences were processed using the LOME FrameNet parser [Xia et al. 2021]. Parser outputs were extracted and manually reviewed to identify errors, including missing annotations and incorrect frame assignments. Only lexical unit labels were retained for analysis, as they directly indicate the frames evoked by adjectives, as indicated in §2.3.

5. Results and Discussion

5.1. Adjective Functions

A total of 182 adjectives were identified in the Brazilian Portuguese descriptions and 165 in the English ones. In both languages, experiential Epithets predominate (85.71% in Portuguese; 81.21% in English), followed by Classifiers, while interpersonal Epithets are rare and occur only in English (1.21%).

This distribution reflects the descriptive nature of image captions, which prioritize observable and visually salient properties. It may also be related to the controlled conditions under which the descriptions were produced, as annotators were asked to describe the images without contextual information (see [Viridiano et al. 2024]. (7) and (8) are descriptions of the same image and illustrate experiential epithets in the corpus:

(7) Um cachorro *preto* e *branco* salta para pegar um brinquedo no gramado de um quintal.
(A black and white dog jumps to get a toy on the lawn of a yard.)

(8) A *black* and *white* dog jumps to get the Frisbee.

The adjectives *preto*, *branco*, *black*, and *white* describe observable physical properties that help distinguish the central entity in the image. Such adjectives are particularly frequent in image descriptions because they provide visually accessible information that is not necessarily specified by the noun itself.

Classifiers occur less frequently because they typically designate stable and well-defined classes of entities. In Flickr30K, however, many of the entities are already basic, visually identifiable participants, such as men, women, children, and animals. As a result, they are more often modified by adjectives expressing qualities, appearance, size, color, age, or evaluation than by classifiers assigning them to specialized taxonomic classes. Examples from our corpus are:

(9) Uma equipe *médica* realiza uma operação.
(A medical team performing an operation.)

(10) *Surgical* team performing surgery.

In (9) and (10), *médica* and *surgical* define an institutional role associated with the noun.

Interpersonal Epithets appear only in English descriptions.

(11) Ballet dancers in a studio practice jumping with *wonderful* form.

In sentence (11), the adjective *wonderful* expresses a subjective evaluation rather than a directly observable property. Unlike experiential Epithets, which tend to describe perceptible attributes, interpersonal Epithets foreground the annotator's stance toward the image.

However, even experiential Epithets involve interpretative variation. Although they are treated as descriptors of observable properties, their use still depends on perceptual categorization. The same visual property may be categorized differently by different annotators, especially in borderline cases such as color distinctions. This indicates that even descriptions that appear to be objective involve choices about how visual information is selected and verbalized.

5.2. Adjectives Across the Two Languages

There were 47 instances of one-to-one adjective correspondences across descriptions of the same image in the two languages. Most correspondences involved experiential Epithets, especially color adjectives, suggesting that visually salient properties tend to be consistently selected by annotators in both languages.

Some examples include:

(12) Criança de vestido *rosa* entrando em uma pequena casa de madeira. *(Child in a pink dress going into a small wooden house.)*

(13) A little girl in a *pink* dress going into a wooden cabin.

The predominance of color adjectives among adjective correspondences suggests that color constitutes one of the most consistently verbalized types of visual information across languages. This tendency reinforces the relevance of visually prominent attributes in image descriptions.

However, equivalence at the level of meaning does not necessarily entail equivalence in grammatical realization. In several cases, meanings realized by adjectives in one language were encoded through other grammatical resources in the other language. From an SFL perspective, this shows that similar experiential meanings may be construed through different configurations of the nominal group or redistributed across other elements of the clause.

For example, in (12) the Portuguese prepositional phrase *de madeira* realizes the same experiential meaning as the English adjective *wooden* in (13). Similar shifts in lexicogrammatical realization occur in descriptions of posture, bodily position, and material composition. These meanings may be construed through different resources across the two languages, including adjectives within the nominal group, verbal processes in the clause, or prepositional phrases functioning as Qualifiers. The following image, Fig. 2, and its descriptions illustrate this tendency:



Figura 2. Image 1000919630 from the Flickr30K dataset)

(14) Um rapaz *sentado* em uma cadeira brinca com um leão *de pelúcia*.
(A guy sitting on a chair plays with a stuffed lion.)

(15) Um jovem *sentado* em um banco e vestindo bermuda, tênis e camiseta segura um leão de pelúcia e mexe na boca do brinquedo.
(A youth sitting on a stool and wearing shorts, sneakers and a T-shirt holds a stuffed lion and touches the toy's mouth.)

(16) A man sits in a chair while holding a large *stuffed* animal of a lion.

(17) A man is sitting on a chair holding a large *stuffed* lion.

These examples show a difference in grammatical realization. The meaning of material or composition is expressed in Portuguese through the prepositional phrase *de pelúcia*, while in English it is realized adjectivally through *stuffed*.

These patterns indicate that cross-linguistic differences in the distribution of adjectives are not only a matter of lexical preference. They also reflect broader differences in how comparable experiential meanings are mapped onto lexicogrammatical resources in each language.

5.3. LOME FrameNet Parser Performance and Limitations

Following the manual annotation of adjective functions and described entities, all sentences were processed using the LOME FrameNet parser [Xia et al. 2021] in order to identify semantic frames evoked by adjectives. The parser generated 445 labels for the 347 adjectives identified in the dataset, including both lexical unit and frame element annotations.

Most labels were correctly assigned, although three main error types were identified during manual revision: missing annotations, incorrect frame assignments, and incorrect frame element assignments. In total, 102 parser errors were identified, including 71 unlabeled adjectives, 26 incorrect frame assignments, and 5 incorrect frame element assignments. Although only 6.97% of the generated labels were incorrect, 20.46% of the adjectives in the dataset were left unlabeled.

One example of incorrect frame assignment occurs in the following sentence:

(18) A bunch of people in racing gear and helmets racing bicycles around a corner on *wet* pavement.

In this case, the adjective *wet* was incorrectly associated with the frame `Emptying` instead of `Being_wet`.

The parser also failed to identify some adjectives entirely. For example, in the sentence:

(19) Two *young* guys with *shaggy* hair look at their hands while hanging out in the yard.

While *young* was correctly associated with the frame `Age`, the adjective *shaggy* received no annotation.

Despite these limitations, the parser proved useful for large-scale semantic annotation, particularly when combined with manual curation. The results also highlight challenges involved in automatic frame-semantic annotation in multimodal contexts, especially in cases where semantic interpretation depends on visual or contextual information not directly available in the textual input.

5.4. Frames evoked by Adjectives

After removing incorrect annotations, only lexical unit labels were considered in the analysis of frames evoked by adjectives, since they directly indicate the semantic frames associated with each lexical unit. A total of 41 distinct frames were identified across the dataset.

The most frequent frame in both languages was `Color` (49.50% in English; 39.54% in Portuguese), reflecting the prominence of color-related information in image descriptions. Other recurrent frames include `Age` and `Size` in English, and `Body movement` and `Change Posture` in Brazilian Portuguese.

The predominance of the `Color` frame is consistent with the high frequency of color adjectives identified throughout the dataset, indicating that color is one of the most consistently verbalized types of visual information across languages.

While the two languages showed relatively similar distributions of adjective functions, frame analysis revealed more substantial semantic differences. Functional categories such as experiential Epithets capture broad grammatical behavior, whereas frame-semantic analysis highlights finer distinctions in the types of meanings associated with adjectives across languages. Portuguese descriptions showed greater frame diversity overall, with 32 distinct frames, compared to 16 in English. English descriptions, in contrast, concentrated a larger proportion of adjectives within a smaller set of semantic domains, particularly `Color` and `Age`. More substantial differences emerge in descriptions involving bodily position and posture. In Brazilian Portuguese, these meanings are frequently encoded through adjectives:

> (20) Uma pessoa, de bermuda *azul*, *deitado* no chão ao lado de carros.
> *(A person, in blue shorts, lying on the ground next to cars.)*
>
> In English, equivalent semantic content is often realized through verbal constructions:
>
> (21) A man in *blue* shorts is *lying* in the street.

In sentence (20), the adjective *deitado* evokes the `Change Posture` frame, whereas in (21) the same semantic information is encoded through the verb *lying*. As a result, certain frames become associated with adjectives in Portuguese, but not in English, despite being conceptually present in both descriptions. These patterns suggest that cross-linguistic differences in frame distribution may be related to semantic, typological, and syntactic differences in how visual information is linguistically encoded.

6. Conclusions

This study investigated the use of adjectives in image descriptions in Brazilian Portuguese and English by combining Systemic Functional Grammar and Frame Semantics in a multimodal analytical framework. Rather than focusing exclusively on grammatical categories, the study examined how adjectival meanings contribute to the semantic representation of visual information across languages.

The results showed that experiential Epithets predominate in both datasets, reflecting the descriptive and perceptually oriented nature of image captions. At the semantic level, the analysis revealed the prominence of visually salient domains such as `Color`,

while also identifying relevant cross-linguistic differences in the distribution of semantic frames.

The findings also demonstrate that semantically equivalent information is not necessarily realized through equivalent grammatical structures across languages. Information expressed adjectivally in one language may be encoded through verbal or prepositional constructions in the other, indicating that cross-linguistic variation in image descriptions involves not only lexical choice, but also broader semantic and typological preferences in how visual experience is linguistically structured.

From a methodological perspective, the study suggests that the combination of functional and frame-semantic analyses provides complementary insights into multimodal language description. While Systemic Functional Grammar captures broad grammatical tendencies, Frame Semantics enables a more fine-grained analysis of the conceptual domains activated by adjectives and their relation to visual interpretation. This distinction becomes particularly relevant in multimodal contexts, where different linguistic strategies may encode similar perceptual information.

The results also reinforce the relevance of frame-based semantic annotation for multimodal resources and accessible image description. Understanding how languages differ in the semantic organization of visual information may contribute to the development of more linguistically sensitive multimodal annotation practices and automatic caption generation systems.

Although exploratory and based on a limited dataset, this study contributes to discussions on cross-linguistic multimodal description, semantic annotation, and the linguistic representation of visual experience. Future research may expand the corpus, include additional image genres, and further investigate how semantic frames interact with translation, accessibility, and multimodal language technologies.

7. Acknowledgments

The authors thank two anonymous reviewers whose comments helped us improve this article. The usual disclaimer applies. Maucha A. Gamonal was funded by CAPES (88887.015648/2024-00) and is currently supported by H.IAAC 01245.003479/2024-10. Adriana Pagano has grants from the National Council for Scientific and Technological Development (CNPq 404722/2024-5; 313103/2021-6) and Minas Gerais State Agency for Research and Development (FAPEMIG). André V. Lopes Coneglian's research is funded by CNPq (153972/2025-4).

8. Limitations

The study is exploratory and has some limitations. The analysis is based on a small sample of 30 images from the Framed Multi30K dataset, resulting in 300 sentences, and the findings should therefore be interpreted as tendencies rather than as statistically generalizable patterns. In addition, the sample does not cover the full range of possible image types, descriptive styles, or annotator variation. The frame-semantic annotation was supported by the LOME parser and manually revised, so the results depend both on the parser's coverage and on interpretive decisions made during revision. Finally, the focus on adjectives necessarily excludes other lexicogrammatical resources through which

similar visual meanings may be realized, such as verbal and prepositional constructions. Future work should expand the sample and examine how visual attributes are distributed across different grammatical resources.

9. Ethics statement

This study does not involve collecting data from human participants or sensitive information of any kind. The Framed Multi30K dataset is openly accessible. All tools used for corpus compilation and annotation, code generation, and data analysis are freely available for research use.

10. Generative AI usage

No scientific content, analyses, or interpretations were produced by AI.

11. References

Referências

Belcavello, F., Torrent, T. T., Matos, E. E., Pagano, A. S., Gamonal, M., Sigiliano, N., Dutra, L. V., de Andrade Abreu, H., Samagaio, M., Carvalho, M., Campos, F., Azalim, G., Mazzei, B., de Oliveira, M. F., Loçasso Luz, A. C., Pádua Ruiz, L., Bellei, J., Pestana, A., Costa, J., Rabelo, I., Silva, A. B., Roza, R., Souza, M., and Oliveira, I. (2024). Frame2: A FrameNet-based multimodal dataset for tackling text-image interactions in video. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 7429–7437, Torino, Italia. ELRA and ICCL.

Coneglian, A. V. L. and Pagano, A. (2025). A gramática em datasets multimodais: um estudo de caso de legendas de imagens. *Caligrama: Revista de Estudos Românicos*, 30(1):24–51.

Dixon, R. M. (2010). *Basic linguistic theory volume 2: Grammatical topics*, volume 2. OUP Oxford.

Fillmore, C. J. (1982). Frame semantics. In Linguistics Society of Korea, editor, *Linguistics in the Morning Calm*, pages 111–138. Hanshin Publishing Co., Seoul, South Korea.

Halliday, M. A. K. and Matthiessen, C. M. I. M. (2013). *Halliday's Introduction to Functional Grammar*. Routledge.

Ruppenhofer, J., Ellsworth, M., Schwarzer-Petruck, M., Johnson, C. R., Baker, C. F., and Scheffczyk, J. (2016). *FrameNet II: Extended Theory and Practice*. International Computer Science Institute, Berkeley, CA.

Torrent, T. T., Matos, E. E. d. S., Belcavello, F., Viridiano, M., Gamonal, M. A., Costa, A. D. d., and Marim, M. C. (2022). Representing context in framenet: A multidimensional, multimodal approach. *Frontiers in Psychology*, 13.

Viridiano, M., Lorenzi, A., Torrent, T. T., Matos, E. E., Pagano, A. S., Sathler Sigiliano, N., Gamonal, M., de Andrade Abreu, H., Vicente Dutra, L., Samagaio, M.,

Carvalho, M., Campos, F., Azalim, G., Mazzei, B., Fonseca de Oliveira, M., Luz, A. C., Padua Ruiz, L., Bellei, J., Pestana, A., Costa, J., Rabelo, I., Silva, A. B., Roza, R., Souza Mota, M., Oliveira, I., and Pelegrino de Freitas, M. H. (2024). Framed Multi30K: A frame-based multimodal-multilingual dataset. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 7438–7449, Torino, Italia. ELRA and ICCL.

Xia, P., Qin, G., Vashishtha, S., Chen, Y., Chen, T., May, C., Harman, C., Rawlins, K., White, A. S., and Van Durme, B. (2021). LOME: Large ontology multilingual extraction. In Gkatzia, D. and Seddah, D., editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 149–159, Online. Association for Computational Linguistics.

Clause Linkage in Brazilian Portuguese Narratives: Evidence from Enhanced Dependency Annotation

Letícia Guedes Guimarães[1], Adriana Pagano[1], André V. Lopes Coneglian[1], Magali Duran[2]

[1] Universidade Federal de Minas Gerais
[2]Universidade de São Paulo
{letigg, apagano, coneglian}@ufmg.br
magali.duran@gmail.com

Abstract. *This paper explores the potential of Enhanced and Extended Enhanced Universal Dependencies for narrative analysis in Brazilian Portuguese, focusing on how enhanced dependency relations contribute to clause linkage. The study examines narrative sentences annotated in CoNLL-U format and compares basic and enhanced dependency layers. The results show that enhanced dependencies are particularly useful for recovering implicit subject relations in Portuguese and for enriching clause relations through information from prepositions and conjunctions. These enhancements enable a more explicit representation of participant continuity and event relations in narratives, supporting the transition from syntactic dependency analysis to richer semantic annotation.*

1. Introduction

Clause linkage is a central phenomenon for the linguistic description of discourse, since it concerns the ways in which clauses are combined, coordinated, embedded, or otherwise connected within larger sentence units [Lehmann 1988, Matthiessen and Thompson 1988]. From a systemic-functional perspective, clause complexes are one of the main grammatical resources through which logico-semantic relations between clauses are organized, including elaboration, extension, enhancement, projection, and different degrees of interdependency [Halliday and Matthiessen 2014]. In narrative texts, such as fables, these resources are especially relevant because they contribute to the sequencing of events, the backgrounding and elaboration of circumstances, the attribution of speech and the tracking of participants across clauses.

From a computational perspective, clause linkage also raises important challenges for syntactic annotation. Basic dependency representations usually encode the local head-dependent relations that hold between words, but they do not always make explicit the full range of syntactic and semantic dependencies involved in complex clause combinations. This is particularly relevant in cases such as coordinated predicates with shared subjects, control constructions, relative clauses, and constructions involving argument sharing across clauses. In such cases, some dependencies are recoverable by human interpretation but remain implicit in the basic syntactic annotation.

Enhanced Universal Dependencies (EUD) [Schuster and Manning 2016] provide a useful framework for addressing these phenomena. By allowing additional dependency relations to be represented beyond the basic tree structure, enhanced dependencies can

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 499–511
October 19-22, 2026 ©2026 Association for Computational Linguistics

make implicit syntactic links more explicit. For example, they can represent shared arguments in coordination, propagated subjects in control constructions, and additional relations involving relative clauses. These enriched representations are therefore relevant not only for linguistic analysis, but also for downstream natural language processing tasks that depend on more explicit predicate-argument structures.

This paper explores the potential of EUD for informing narrative analysis in Brazilian Portuguese. The study examines how enhanced dependency annotation can make explicit relations relevant to narrative interpretation, with a focus on clause linkage and related phenomena, such as participant continuity, reported speech, relative-clause coreference, and the semantic specification of coordinating and subordinating relations. The analysis is intended as an intermediate step in a broader annotation pipeline. By examining which discourse-relevant relations can be recovered through EUD, we identify correspondences between enhanced dependency structures and phenomena represented more explicitly in subsequent semantic annotation, particularly in UMR [Van Gysel et al. 2021], such as coreference, temporal relations, modal dependencies, and event structure. In this sense, the paper situates enhanced dependency annotation as an intermediate layer between syntactic dependency analysis and richer semantic representations of narrative discourse.

2. Related work

In contrast to constituency-based representations, which organize words into hierarchically nested phrases, basic dependency trees, such as those used in the Universal Dependencies framework [Nivre et al. 2020], represent morphosyntactic structure through labeled binary relations between individual words. For example, the relation *nsubj(arrancou, cigarra)* encodes a direct dependency between the verb *arrancou* and its nominal subject, *cigarra*. A constituency analysis would represent the same grammatical relation indirectly, through the position of the noun phrase containing *cigarra* within the phrase structure. Dependency representations are widely used in natural language processing, particularly in multilingual research, because they provide a common framework for describing grammatical relations across languages with different word-order patterns. However, basic dependency trees have long been recognized as limited in their ability to represent semantically relevant relations between content words when these relations are not directly expressed in surface syntactic structure [Schuster and Manning 2016].

This limitation was progressively addressed through richer annotation schemes: first with the collapsed relations of the Stanford Dependencies framework [de Marneffe et al. 2006, de Marneffe and Manning 2008], which expanded the representation of prepositional and conjunctive relations, and later with Enhanced Universal Dependencies (EUD), proposed by [Schuster and Manning 2016], which extended this approach to a broader range of grammatical phenomena within the UD framework, such as shared arguments, elided subjects, and dependencies in relative clauses.

According to the UD guidelines [Nivre et al. 2020, de Marneffe et al. 2021], EUD complement basic dependencies by making implicit information explicit and by enriching dependency labels to better disambiguate arguments and modifiers. This can be implemented either through the addition of new relations or through the modification of existing ones. These two processes differ not only in what they represent linguistically, but also

in how amenable they are to automatic annotation: additive relations tend to be more easily converted automatically, while modification-based ones generally demand human revision [Pagano et al. 2023].

EUD was first introduced in Brazilian computational linguistics by [Pagano et al. 2023], providing examples in Portuguese for each type of enhanced dependency and discussing strategies for implement them. A detailed study was then elaborated by [Duran 2024] and [Duran et al. 2024], defining several rules for enhanced dependencies in Portuguese, including the propagation of subjects elided in pro-drop languages. These rules were used to customize a converter from UD to EUD for Portuguese [Souza et al. 2024]. This converter was improved one year later with rules developed for pro-drop languages.These rules, referred to as Extended Enhanced Universal Dependencies or EEUD, extended the original framework, introducing additional relations to represent implicit arguments, particularly omitted subjects in languages where verbal morphology allows their recovery. The converter was used to produce a layer with both EUD and EEUD in the corpus Porttinari-base [Duran et al. 2025], the first corpus in Portuguese annotated with enhanced dependencies. Together, these efforts laid the groundwork for the annotation reported in this paper.

EUD annotation is especially relevant to the analysis of narrative texts because enhanced dependencies capture relations that are more explicit and semantically informative than those represented in basic dependency structures, thereby offering useful insights into narrative organization and interpretation. Despite this potential, the application of EUD to narrative analysis remains largely unexplored, a gap that the present study seeks to address. Fables are particularly suitable for this purpose because they typically combine two linguistically distinct components—a narrative sequence and a moral commentary—creating opportunities for future comparisons of the relations instantiated in each. From a computational perspective, Brazilian Portuguese narrative datasets annotated according to internationally established frameworks such as UD and EUD also remain scarce, since existing UD resources for Brazilian Portuguese predominantly represent other registers. By providing EUD annotation for a corpus of fables, this study contributes to filling this resource gap and broadens the possibilities for computational and linguistic analyses of narrative discourse.

3. Methodology

3.1. Corpus compilation and annotation

Our corpus is part of a larger multilingual corpus of Aesop's fables translated from Greek into Brazilian Portuguese, French, English, Spanish, Russian, Chinese and Italian. The Brazilian Portuguese translations were published in *Esopo: fábulas completas* [Dezotti et al. 2013] and are available in print only. The text was manually typed and then saved in two separate plain-text files, one compiling the narrative section of the fables and another one compiling the moral or epimythium section [Dezotti 1988].

The full corpus is outlined in Table 1. In this paper, we report on the narrative subset, which accounts for 85% of the total corpus.

Table 1. Corpus composition and basic statistics.

Corpus	Sentences	Tokens	Unique words	Lemmas	Avg. sentence length
Narratives	355	8,836	2,163	1,679	24.89
Epimythia	63	1,486	512	444	23.59
Full corpus	418	10,322	2,440	1,857	24.69

3.1.1. Basic dependencies annotation

The corpus was automatically annotated for basic dependencies using the online parser UD-Pipe [Straka et al. 2016] trained on the PetroGold corpus [Souza et al. 2021]. The resulting CoNLL-U files were uploaded to Arborator-Grew [Guibon et al. 2020] for manual revision. Tokenization errors involving Brazilian Portuguese contractions that failed to be properly detokenized (e.g., 'comigo' instead of 'com mim', 'dele' instead of 'de ele') due to adjacent punctuation marks were corrected before the revision of the basic UD annotation, which was conducted in two rounds: an initial pass by one language specialist, followed by adjudication by two qualified reviewers.

Revision followed the Brazilian Portuguese UD annotation guidelines [Duran 2021, Duran 2022]. During this stage, the automatic parses were corrected and supplemented with dependency subrelations. For instance, reported-speech clauses were manually annotated with the `ccomp:speech` subrelation.

3.2. Enhanced dependencies annotation

Enhanced annotation was carried out using a converter customized for Portuguese by [Souza et al. 2024], which includes, in its last version, all EEUD proposed by [Duran et al. 2025]. A description of each category and examples can be found in [Pagano et al. 2023] and [Duran et al. 2025].

The repository was cloned locally and the scripts were run in Visual Studio Code over the CoNLL-U files used in this study. The output consists of CoNLL-U files in which the basic dependency layer is preserved and the enhanced dependency graph is added in the `DEPS` column. Manual revision was carried out on all sentences using a customized UD annotation and visualization tool [Souza 2026]. The quality and limitations of the automatic conversion exceed the scope of this paper and will be evaluated in future work.

3.3. Data extraction and analysis

Quantitative data extraction was performed using a customized Python script in Visual Studio Code. The script was designed to extract corpus-level statistics, part-of-speech frequencies, basic dependency relations, and enhanced dependency relations from the CoNLL-U files. It also compared the basic `HEAD:DEPREL` pairs with the relations encoded in the enhanced dependency column in order to identify how the enhanced layer contributes additional syntactic and semantic information.

The analysis distinguishes between enhanced relations that add new dependency links and enhanced relations that specify existing basic links. Added links include, for example, propagated subjects, propagated conjunct dependencies, relative-clause coreference links, and reattached complements. By contrast, relations such as `obl:de`, `nmod:de`, `advcl:quando`, and `conj:e` usually preserve the same head as the corresponding basic relation and add lexical or semantic information after the colon. These

were therefore treated as enhanced specifications of basic relations, rather than as newly added links.

4. Results

This section presents and discusses the results obtained from the annotated corpus. We first describe the distribution of part-of-speech tags, before examining basic and enhanced dependency relations relevant to clause linkage and narrative interpretation, with a focus on the contributions of enhanced representations.

4.1. Part-of-speech distribution

We first examine the distribution of part-of-speech tags in the corpus. As shown in Table 2, the corpus contains a high proportion of nominal and verbal tags. The most frequent lexical categories are NOUN, with 1,292 occurrences, and VERB, with 1,222 occurrences. Function words are also frequent, especially DET with 1,115 occurrences, ADP with 879 occurrences, and PRON with 762 occurrences. This distribution is consistent with the narrative character of the corpus. Nouns contribute to the introduction and tracking of participants, verbs encode events and actions, and function words contribute to the grammatical organization of clauses and clause complexes. The frequencies of CCONJ (374) and SCONJ (278) are particularly relevant for the present study, since they point to the importance of coordination and subordination in the corpus.

Table 2. UPOS frequency distribution in the corpus.

UPOS	Frequency
PUNCT	1,327
NOUN	1,292
VERB	1,222
DET	1,115
ADP	879
PRON	762
ADV	384
CCONJ	374
SCONJ	278
ADJ	253
AUX	247
PROPN	31
NUM	20
INTJ	3

The relatively low number of proper nouns is also noteworthy. The corpus contains only 31 PROPN tokens. As shown in Table 3, these correspond to a small set of mythological figures and geographical names.

Table 3. Proper nouns in the corpus.

PROPN form	Frequency
Zeus	15
Hermes	8
Prometeu	3
Pireu	3
Súnion	1
Ática	1
Total	31

The low frequency of PROPN tokens indicates that named entities play only a limited role in the corpus. This is consistent with the type of narratives analyzed here, where participants are often referred to by common nouns, such as animal names and social roles, rather than by individual proper names. For this reason, the analysis of participant reference cannot be based primarily on named entities. Instead, it requires attention to how participants are introduced and tracked through nominal expressions, pronouns, and syntactic relations across clauses.

4.2. Basic dependency relations

The basic dependency layer provides the first level of syntactic representation for the corpus. Since the corpus contains 355 sentences, the basic annotation includes 355 `root` relations. Table 4 shows the most frequent basic dependency relations in the corpus.

Table 4. Most frequent basic dependency relations.

Basic dependency relation	Frequency
punct	1,327
det	1,083
case	711
nsubj	558
obj	513
obl	424
mark	410
cc	378
advmod	361
root	355
advcl	333
conj	293
nmod	188
xcomp	163
amod	129

Excluding punctuation, the most frequent relations are `det`, `case`, `nsubj`, `obj`, `obl`, `mark`, and `cc`. These relations reflect the high frequency of nominal modification, prepositional marking, participant encoding, and clause-linking material in the corpus. The presence of frequent `advcl` and `conj` relations is particularly relevant for this study, since these relations directly contribute to the syntactic representation of subordination and coordination.

Relations associated with clause linkage and narrative interpretation in the corpus can be seen in Table 5.

Table 5. Basic dependency relations relevant to clause linkage.

Basic dependency relation	Frequency
advcl	333
conj	293
xcomp	163
ccomp	93
acl:relcl	88
ccomp:speech	64
acl	61
csubj	15
parataxis	10

The basic layer therefore already captures a substantial number of clause-linking

configurations. The high number of advcl relations indicates frequent adverbial subordination, while conj relations point to recurrent coordination between predicates or larger syntactic units. Complementation is represented by both ccomp and xcomp, and the distinction between ccomp and ccomp:speech makes it possible to separate general clausal complements from reported speech complements. Relative and adnominal clauses are represented by acl:relcl and acl, respectively.

Subject-type relations are particularly important for the analysis of participant continuity in narratives. In the basic layer, there are 588 subject-type relations, including 558 nsubj, 15 nsubj:pass, and 15 csubj relations. Object-type relations are also frequent in the basic layer. In the basic dependency layer, object-type relations are represented mainly by obj, with 513 occurrences, and, less frequently, by iobj, with 55 occurrences. Together, these account for 568 core object-type relations.

Taken together, the basic dependency results show that the corpus contains a high number of syntactic structures relevant to clause linkage. Relations such as advcl, conj, xcomp, ccomp, ccomp:speech, acl:relcl, and acl provide the basic syntactic representation of subordination, coordination, complementation, reported speech, and adnominal modification. The following subsection examines how the enhanced dependency layer adds further relations that make participant continuity and clause-linking meanings more explicit.

4.3. Enhanced dependencies

We examined enhanced relations by distinguishing newly introduced links from specifications of existing basic dependencies. The former make implicit syntactic relations explicit and support the analysis of participant tracking and event structure, whereas the latter add semantic detail relevant to clause linkage, including information contributed by prepositions and subordinating or coordinating conjunctions.

Table 6 reports the most frequent enhanced relations observed in the corpus.

Table 6. Most frequent enhanced dependency relations.

Enhanced relation	Frequency
nsubj	591
conj:e	254
nmod:de	169
obl:em	107
ref	86
obl:de	76
advcl:quando	71
obl:com	50
obl:por	50
mark	50
obl:a	48
advcl:se	33
obl:para	32
advcl:a_o	29
obl:arg:a	29
obj	28
advcl:para	24
advcl:pois	23
ccomp	22

4.3.1. Enhanced dependencies and participant tracking

The enhanced layer adds 610 subject-type relations, slightly more than the 588 subject-type relations present in the basic layer. This increase reflects the pro-drop nature of Brazilian Portuguese, elided subjects being frequently recoverable from context. By recovering implicit subject dependencies and propagating subject links across coordinated, subordinate, and non-finite clauses, EEUD makes participant–event relations more explicit. The most frequent added enhanced relation is `nsubj`, with 591 instances, indicating that subject continuity is a major contribution of the enhanced layer. These additional links are particularly relevant in clause-linkage configurations, such as coordinated predicates and control constructions, where the same participant is associated with multiple events. Table 7 compares subject-type relations in basic and enhanced dependencies.

Table 7. Comparison of subject-type relations in the basic and enhanced layers.

Subject-type relation	Basic	Enhanced
`nsubj`	558	591
`nsubj:pass`	15	13
`csubj`	15	6
Total	588	610

Figure 1 illustrates a case of enhanced subject relations, which indicates that the enhanced layer frequently adds participant–predicate links that are not explicit in the basic tree[1].

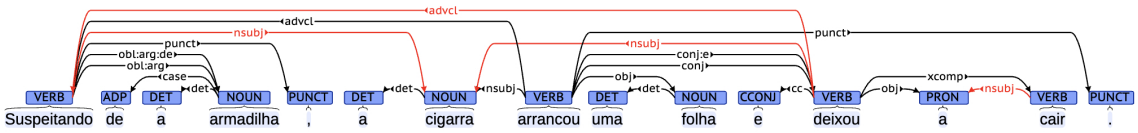


Figure 1. Basic and enhanced dependencies in a sample sentence.

In the basic layer, *a cigarra* is the subject of the finite verb *arrancou*; the gerund-dive clause headed by *Suspeitando* is attached as `advcl`; and *deixou* is coordinated with *arrancou* through `conj`. The enhanced layer makes additional relations explicit: *a cigarra* is also linked as subject to *Suspeitando* and *deixou*; the adverbial clause relation is propagated to the coordinated predicate; and the coordinating conjunction is specified in `conj:e`. In addition, the oblique argument *da armadilha* is specified as `obl:arg:de`, making the prepositional marking explicit. The example also illustrates the assignment of an external subject to an `xcomp`. In the basic layer, the clitic pronoun *a* is annotated as the object of *deixou*. In the enhanced layer, it is also linked as `nsubj` of the infinitive *cair*. This makes explicit that the entity undergoing the falling event is the leaf, resumed by the clitic *a*, rather than the matrix subject *a cigarra*. Thus, the enhanced relation `14:nsubj` avoids an incorrect interpretation in which the subject of the matrix clause would be propagated to the `xcomp`.

[1]In the figure, red arcs highlight enhanced dependencies that add new links to the graph. Black arcs include the basic dependency structure and enhanced relations that duplicate or specify existing basic links.

Unlike subject relations, enhanced dependencies add relatively few object relations. The basic layer contains 568 object-type relations, while the enhanced layer adds only 28 `obj` relations and no additional `iobj` relations.

Table 8. Comparison of object-type relations in the basic and enhanced layers.

Object-type relation	Basic	Enhanced
`obj`	513	28
`iobj`	55	0
Total	568	28

This contrast suggests that, in this corpus, enhanced dependencies contribute more strongly to the recovery of subject continuity than to the recovery of object relations. Added subject links help identify which participant is associated with each predicate, especially in coordinated, subordinate, and non-finite clauses where the participant may be shared but not repeated.

4.3.2. Enhanced dependencies and clause linkage

Enhanced relations also provide finer-grained information about clause linkage. Coordination is particularly prominent. Table 9 shows coordination relations specified in the enhanced layer. These specifications make the semantic or discourse value of coordination more explicit, distinguishing additive, contrastive, comparative, and alternative links. The most frequent coordination relation is additive, linking events or propositions through conjunction. Its frequency may reflect the relatively linear progression of fable narratives, where successive actions and developments are commonly connected through simple additive relations. However, this distribution should be interpreted with caution, since the conjunction `e` can also introduce relations with additional semantic values, such as contrast or temporal succession, depending on the context.

Table 9. Enhanced coordination relations.

Added coordination relation	Frequency
`conj:e`	254
`conj:mas`	19
`conj:como`	2
`conj:ou`	1
Total	276

Adverbial clause relations are also highly informative. As Table 10 shows, the enhanced layer specified relations which allow for an analysis of the semantic relations construed by adverbial clauses, such as temporal sequencing, conditionality, purpose, cause, comparison, and concession. These relations are especially relevant for narrative interpretation because they help identify how events are temporally ordered, causally motivated, conditionally framed, or contrastively related.

Temporal relations such as `advcl:quando`, `advcl:enquanto`, `advcl:assim_que`, `advcl:depois_que`, and `advcl:logo_que`, accounting for most adverbial relations in the corpus, contribute to event sequencing, which is central to narrative texts. Conditional relations such as `advcl:se` help represent

hypothetical or contingent developments. Causal and explanatory relations such as advcl:pois and advcl:porque contribute to the interpretation of motivation and consequence.

Table 10. Selected enhanced adverbial clause relations.

Selected added adverbial clause relation	Frequency
advcl:quando	71
advcl:se	33
advcl:a_o	29
advcl:para	24
advcl:pois	23
advcl:enquanto	19
advcl	19
advcl:como	14
advcl:a	10
advcl:porque	10
advcl:que	9
advcl:assim_que	9
advcl:por	7
advcl:sem	6
advcl:a_o_passo_que	6
advcl:até_que	4
advcl:para_que	4
advcl:depois_que	3
advcl:de_o_que	3
advcl:após	3

Complement clauses also appear among enhanced relations as seen in Table 11. These relations are relevant for representing speech, thought, perception, intention, and control structures, all of which are important in narrative texts.

Table 11. Enhanced complement relations.

Added complement relation	Frequency
ccomp	22
ccomp:speech	12
xcomp	12
Total	46

5. Conclusion

This paper explored the potential of EUD, specifically EEU, for informing the analysis of clause linkage in Brazilian Portuguese narratives. Overall, the enhanced layer contributes information that is implicit in the basic dependency representation. In this corpus, its most important contribution concerns subject continuity, showing that EEUD are valuable for recovering participant–event relations in a pro-drop language such as Brazilian Portuguese. By making implicit relations more explicit and enriching existing dependencies with additional linguistic information, they provide a more detailed account of clause linkage and narrative structure.

Future work will extend this analysis toward UMR annotation [Van Gysel et al. 2021], with particular attention to coreference, temporal relations, modality, and event structure. In this broader pipeline, EEUD can serve as an intermediate layer between syntactic dependency annotation and richer semantic representations of narrative discourse. Upon completion, the corpus will be made available through a GitHub repository.

6. Acknowledgments

The authors thank the anonymous reviewers for their comments, which helped improve this paper. Letícia Guedes is currently supported by Minas Gerais State Agency for Research and Development (FAPEMIG). Adriana Pagano has grants from FAPEMIG and the National Council for Scientific and Technological Development (CNPq 404722/2024-5; 313103/2021-6). André V. Lopes Coneglian's research is funded by CNPq (153972/2025-4). Their work is supported by the National Institute of Science and Technology in Responsible Artificial Intelligence for Computational Linguistics, Information Treatment, and Dissemination (INCT-TILDIAR), funded by CNPq (grant no. 408490/2024-1). Magali Sanches Duran is a member of the Artificial Intelligence Center of the University of São Paulo, funded by the São Paulo Research Foundation (FAPESP grant no. 2019/07665-4) and IBM. She is also supported by the Ministry of Science, Technology and Innovations, with resources from Law No. 8,248 of October 23, 1991, within the scope of PPI-Softex, coordinated by Softex and published as ICT Residency 13, DOU 01245.010222/2022-44.

7. Limitations

The study has some limitations. The corpus is relatively small and restricted to narrative sentences, and the analysis is primarily descriptive and annotation-oriented. In addition, the enhanced dependencies were produced by rule-based procedures, which means that some cases require manual validation, especially where enhanced relations depend on contextual interpretation. The impact of different automatic parsers on the efficiency of manual revision and on the quality of the resulting enhanced annotations remains to be investigated.

8. Ethics statement

This study does not involve collecting data from human participants or sensitive information of any kind. The corpus consists of translated texts made available for research purposes with the translator's explicit permission and in compliance with copyright requirements. All tools used for corpus compilation and annotation, code generation, and data analysis are freely available for research use.

9. Generative AI usage

No scientific content, analyses, or interpretations were produced by AI.

References

de Marneffe, M.-C., MacCartney, B., and Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In Calzolari, N., Choukri, K., Gangemi, A., Maegaard, B., Mariani, J., Odijk, J., and Tapias, D., editors, *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

de Marneffe, M.-C. and Manning, C. D. (2008). The Stanford typed dependencies representation. In Bos, J., Briscoe, E., Cahill, A., Carroll, J., Clark, S., Copestake, A., Flickinger, D., van Genabith, J., Hockenmaier, J., Joshi, A., Kaplan, R., King, T. H.,

Kuebler, S., Lin, D., Lønning, J. T., Manning, C., Miyao, Y., Nivre, J., Oepen, S., Sagae, K., Xue, N., and Zhang, Y., editors, *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8, Manchester, UK. Coling 2008 Organizing Committee.

de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal dependencies. *Computational Linguistics*, 47(2):255–308.

Dezotti, M. C. C. (1988). A fábula esópica anônima: Uma contribuição ao estudo dos "atos de fábula". Master's thesis, Universidade Estadual Paulista, Araraquara, Brazil.

Dezotti, M. C. C., Berliner, E., and Duarte, A. (2013). *Esopo: fábulas completas*. Cosac Naify, São Paulo, Brazil.

Duran, M. (2021). Manual de anotação de pos tags: orientações para anotação de etiquetas morfossintáticas em língua portuguesa, segundo as diretrizes da abordagem universal dependencies. Technical Report Relatório Técnico No. 434, Universidade de São Paulo, ICMC. Accessed: 2026-04-10.

Duran, M. (2022). Manual de anotação de relações de dependência: versão revisada e estendida. Technical Report Relatório Técnico No. 440, Universidade de São Paulo, ICMC. Accessed: 2026-04-30.

Duran, M., Souza, E., das Graças Volpe Nunes, M., Pagano, A. S., and Pardo, T. A. S. (2025). Extending the enhanced universal dependencies: Addressing subjects in pro-drop languages. In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 143–152, Ljubljana, Slovenia. Association for Computational Linguistics.

Duran, M. S. (2024). Anotação de enhanced dependencies. Available at: `https://repositorio.usp.br/directbitstream/ebb807a6-8543-4616-adb9-2f1715236733/3209188.pdf`. Accessed: 31 May 2026.

Duran, M. S., Nunes, M. d. G. V., and Pardo, T. A. S. (2024). Inferências baseadas em sintaxe: a anotação de sujeitos implícitos. In Claro, D. B. and Pagano, A., editors, *Proceedings of the 15th Brazilian Symposium in Information and Human Language Technology*, pages 356–364, Belém do Pará, Brazil. Association for Computational Linguistics.

Guibon, G. et al. (2020). When collaborative treebank curation meets graph grammars. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC)*, pages 5291–5300.

Halliday, M. A. K. and Matthiessen, C. M. I. M. (2014). *Halliday's Introduction to Functional Grammar*. Routledge, London, 4 edition.

Lehmann, C. (1988). Towards a typology of clause linkage. *Studies in Language*, 12(2):181–225.

Matthiessen, C. and Thompson, S. A. (1988). The structure of discourse and "subordination". In Haiman, J. and Thompson, S. A., editors, *Clause Combining in Grammar and Discourse*, pages 275–329. John Benjamins, Amsterdam.

Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., and Zeman, D. (2020). Universal Dependencies v2: An evergrowing multilingual treebank collection. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Pagano, A. S., Duran, M. S., and Pardo, T. A. S. (2023). Enhanced dependencies para o português brasileiro. In Pardo, T. A. S., Duran, M. S., and Lopes, L., editors, *Proceedings of the 2nd Edition of the Universal Dependencies Brazilian Festival*, pages 461–470, Belo Horizonte, Brazil. Association for Computational Linguistics.

Schuster, S. and Manning, C. D. (2016). Enhanced english universal dependencies: An improved representation for natural language understanding tasks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*.

Souza, E. (2026). Anotador de dependências sintáticas enriquecidas (eud) para o português. Web interface and source code. Source code available at `https://github.com/alvelvis/eud-portugues`. Accessed: 31 May 2026.

Souza, E., Duran, M., Nunes, M. d. G. V., Sampaio, G., Belasco, G., and Pardo, T. (2024). Automatic annotation of enhanced Universal Dependencies for Brazilian Portuguese. In Claro, D. B. and Pagano, A., editors, *Proceedings of the 15th Brazilian Symposium in Information and Human Language Technology*, pages 22–31, Belém do Pará, Brazil. Association for Computational Linguistics.

Souza, E., Silveira, A., Cavalcanti, T., Castro, M., and Freitas, C. (2021). PetroGold – corpus padrão ouro para o dominio do petroleo. In Ruiz, E. E. S. and Torrent, T. T., editors, *Proceedings of the 13th Brazilian Symposium in Information and Human Language Technology*, pages 29–38, Porto Alegre, Brazil. Association for Computational Linguistics.

Straka, M., Hajič, J., and Straková, J. (2016). Udpipe: Trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, pages 4290–4297.

Van Gysel, J. E., Vigus, M., Chun, J., Lai, K., Moeller, S., Yao, J., O'Gorman, T., Cowell, A., Croft, W., Huang, C.-R., et al. (2021). Designing a uniform meaning representation for natural language processing. *KI-Künstliche Intelligenz*, 35(3):343–360.

Inferência Textual *Zero-Shot* em Português: Especialização Linguística *vs.* Raciocínio Explícito em Modelos de Linguagem Compactos

Gabriel Silvestre Mancini[1] **, Marcos Lopes**[1,2]

[1]Programa de Educação Continuada em Engenharia (PECE)
Escola Politécnica – Universidade de São Paulo (USP)

[2]Departamento de Linguística
FFLCH – Universidade de São Paulo (USP)

gabrielsmancini@gmail.com, marcoslopes@usp.br

Abstract. *This work investigates the inferential capabilities of small and medium-sized language models (up to 9B parameters) on Natural Language Inference (NLI) in Brazilian Portuguese under a strictly zero-shot regime. Grounded in the formal semantic definition of entailment as truth-condition preservation, a deductive relation in Peirce's tripartite classification [Peirce 1932], we compare Portuguese-specialized models (Sabiá, Bode, Gaia) with recent multilingual models that incorporate explicit reasoning and knowledge distillation (Qwen3, Gemma 2, Ministral 3) on ASSIN 2 and on a controlled synthetic dataset grounded in lexical inclusion (hyponymy/hyperonymy). Results show that explicit reasoning and distillation are more decisive than monolingual adaptation alone, and that high accuracy on ASSIN 2 may overestimate inferential competence by conflating inductive distributional similarity with deductive truth-condition preservation.*

Resumo. *Este trabalho investiga a capacidade inferencial de modelos de linguagem de pequeno e médio porte (até 9B de parâmetros) na tarefa de Inferência em Linguagem Natural (NLI) em português brasileiro, em regime estritamente* zero-shot*. Partindo da definição formal de acarretamento como preservação de condições de verdade, uma relação dedutiva na classificação peirceana [Peirce 1932], comparamos modelos especializados em português (Sabiá, Bode, Gaia) com modelos multilíngues equipados com mecanismos explícitos de raciocínio e destilação de conhecimento (Qwen3, Gemma 2, Ministral 3) no ASSIN 2 e em um* dataset *sintético controlado, fundamentado em acarretamento por inclusão lexical (hiponímia/hiperonímia). Os resultados indicam que raciocínio explícito e destilação são mais decisivos do que adaptação monolíngue isolada, e que acurácia elevada no ASSIN 2 pode superestimar a competência inferencial ao confundir similaridade distribucional indutiva com preservação dedutiva de condições de verdade.*

1. Introdução

Na semântica formal, o acarretamento é definido como uma relação de preservação de condições de verdade entre sentenças: S_1 acarreta S_2 quando, nos casos em que S_1 é verdadeira, S_2 também é. Essa noção lógica estrita, que na tipologia peirceana corresponde à

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 512–521
October 19-22, 2026 ©2026 Association for Computational Linguistics

dedução, o único tipo de inferência que garantidamente preserva a verdade [Peirce 1932], distingue-se radicalmente de outros tipos de relação entre sentenças, como a similaridade semântica distribucional, aferida *indutivamente* por graus de semelhança léxico-contextual de padrões observados em dados de treinamento.

É precisamente essa dissociação entre relação dedutiva (acarretamento) e reconhecimento indutivo (similaridade) que torna o Reconhecimento de Inferência Textual (NLI) um desafio em aberto, especialmente para modelos compactos (até 9B de parâmetros) em regime estritamente *zero-shot*. Dois paradigmas concorrentes disputam esse espaço: modelos *especialistas* em português, via pré-treinamento contínuo ou ajuste fino instrucional (Sabiá [Pires et al. 2023], Bode [Garcia et al. 2024], Gaia [Camilo-Junior et al. 2025]), e modelos *generalistas* multilíngues com mecanismos explícitos de raciocínio e destilação de conhecimento (Qwen3 [Yang et al. 2025], Ministral 3 [Liu et al. 2026], Gemma [Gemma Team 2024]).

Nesse cenário, o presente trabalho busca contribuir em duas frentes. Primeiro, avalia comparativamente onze LLMs no *benchmark* ASSIN 2 [Real et al. 2020] em regime *zero-shot*. Segundo, introduz um *dataset* sintético controlado, fundamentado em acarretamento por inclusão lexical (hiponímia/hiperonímia) e alinhado à definição formal mencionada anteriormente, para mitigar efeitos de familiaridade prévia dos modelos com o *benchmark* e isolar a competência inferencial genuína.

2. Fundamentação e Trabalhos Relacionados

2.1. Inferência lógica: dedução, indução e abdução

Peirce classifica as inferências lógicas em três tipos com estruturas de raciocínio distintas [Peirce 1932, Peirce 1934]. A **dedução** parte de uma regra geral e de um caso para uma conclusão necessariamente verdadeira. É o único tipo em que a verdade das premissas *garante* a verdade da conclusão. A **indução** generaliza a partir de casos observados para uma regra probabilística: opera ampliativamente e está sujeita a revisão pela experiência. A **abdução** gera hipóteses explicativas para fatos surpreendentes, produzindo conclusões apenas prováveis.

Essa tipologia é diretamente relevante para a tarefa de NLI. O acarretamento semântico, relação que NLI pretende reconhecer, é por definição dedutivo: $S_1 \Rightarrow S_2$ significa que não existe situação em que S_1 seja verdadeira e S_2 falsa. Modelos de linguagem, porém, são treinados por estimativa de máxima verossimilhança, um processo fundamentalmente *indutivo*: aprendem regularidades estatísticas dos dados de treinamento e generalizam a partir delas. A questão empírica central deste trabalho é, portanto, se o treinamento indutivo, com ou sem especialização em português, é capaz de produzir comportamento dedutivo robusto na verificação de condições de verdade.

2.2. Acarretamento semântico e inclusão lexical

Na semântica formal, a operacionalização linguística mais direta do acarretamento é a *inclusão de sentidos*: se um predicado contém um hipônimo, a hipótese obtida pela substituição desse hipônimo por seu hiperônimo é necessariamente acarretada. Por exemplo, *O lateral-esquerdo sofreu a falta* acarreta *O jogador sofreu a falta* porque o conjunto de situações em que um lateral-esquerdo sofre uma falta está contido no conjunto de

situações em que um jogador sofre uma falta. Essa relação é estritamente semântica: independe de implicaturas conversacionais, pressuposições ou conhecimento enciclopédico extralinguístico.

A distinção entre acarretamento (relação lógica, dedutiva, baseada em condições de verdade) e similaridade semântica distribucional (grau de semelhança léxico-contextual, inferida indutivamente) é central para interpretar tanto o *design* dos *benchmarks* quanto o comportamento dos modelos avaliados.

2.3. Benchmarks de NLI em português

Os principais recursos para NLI em português são o SICK-BR [Real et al. 2018], adaptação do corpus SICK do SemEval-2014 para o português brasileiro, e o ASSIN 2 [Real et al. 2020]. O ASSIN 2 representa um avanço metodológico importante ao adotar classificação binária (acarretamento *vs.* não-acarretamento) com sentenças simples extraídas do SICK-BR, reduzindo complexidades sintáticas e discursivas que dificultavam a avaliação controlada do raciocínio semântico.

No espectro supervisionado, o BERTimbau [Souza et al. 2020] com *fine-tuning* no ASSIN 2 estabeleceu o estado da arte com acurácia próxima a 0,90. Em contraste, [da Silva et al. 2023] demonstraram que heurísticas simples, sobreposição lexical, detecção de negação e diferença de comprimento entre premissa e hipótese, atingem F1 $\approx$ 0,71, evidenciando vieses estruturais no *benchmark* que permitem desempenho competitivo sem inferência semântica profunda. Esse resultado levanta a hipótese de que parte do desempenho dos modelos supervisados decorre de exploração de regularidades do *dataset* e não necessariamente de modelagem explícita de inferência lógica. O intervalo entre o piso heurístico (0,71) e o teto supervisionado (0,90) define o espaço no qual posicionamos a avaliação dos LLMs *zero-shot*.

2.4. Modelos compactos: especialização *vs.* raciocínio

Duas vertentes recentes competem no campo dos LLMs compactos. A primeira aposta em **especialização linguística**: o Sabiá [Pires et al. 2023] aplica pré-treinamento contínuo de LLaMA sobre corpus em português (ClueWeb 22); o Bode [Garcia et al. 2024] usa ajuste fino com LoRA em instruções traduzidas (Alpaca); o Gaia [Camilo-Junior et al. 2025] combina pré-treinamento contínuo (13B tokens em português) com *weight merging* para restaurar capacidade instrucional.

A segunda aposta em **raciocínio explícito e destilação**: o Qwen3 [Yang et al. 2025] integra *thinking mode* com orçamento computacional adaptativo (*Chain-of-Thought*); o Ministral 3 [Liu et al. 2026] usa *Cascade Distillation* e pós-treinamento com GRPO em cadeias de pensamento; o Gemma 2 [Gemma Team 2024] substitui a previsão de tokens por destilação de conhecimento a partir de modelos maiores (27B), enriquecendo representações inferenciais.

3. Metodologia

3.1. Datasets

ASSIN 2. O ASSIN 2 [Real et al. 2020] foi carregado via Hugging Face e pré-processado para eliminação de pares duplicados exatos: pares com mesma concatenação

premissa——hipótese em mais de uma partição foram removidos para evitar inflação artificial de métricas. O conjunto resultante contém 9.304 instâncias (4.670 não-acarretamento, 4.634 acarretamento).

Dataset sintético. Como instrumento complementar de validação, foi construído um *dataset* sintético de 60 pares de sentenças, todos no mesmo domínio lexical (futebol), fundamentado exclusivamente na definição formal de acarretamento por inclusão hiponímica. Nos 30 casos de acarretamento, a hipótese resulta da substituição de um hipônimo da premissa por seu hiperônimo (por exemplo, *lateral-esquerdo* $\rightarrow$ *jogador*); nos 30 casos de não-acarretamento, a relação hipônimo-hiperônimo é violada ou invertida. Os pares foram sinteticamente gerados por GPT-5.2 a partir de *prompt* restritivo e submetidos a revisão manual sistemática e em pares, que verificou: (i) a correção gramatical das sentenças; (ii) a validade da relação hiponímica entre os termos substituídos; (iii) a ausência de outros mecanismos inferenciais além da inclusão lexical (como negação, pressuposição ou implicatura); e (iv) a adequação dos casos de não-acarretamento, assegurando que a hipótese não decorra da premissa por nenhum outro mecanismo semântico formal. O modelo usado na geração não integra o conjunto avaliado. A inclusão desse conjunto justifica-se pela possibilidade de que modelos avaliados tenham tido acesso ao ASSIN 2 durante o treinamento, produzindo vieses de familiaridade. Por instrução explícita, o dataset sintético isola o comportamento dos modelos em inferências baseadas em inclusão lexical, sem regularidades superficiais do *benchmark*.

3.2. Avaliação

A avaliação é estritamente *zero-shot*: nenhum modelo foi ajustado, adaptado ou exposto a exemplos da tarefa. Cada par *premissa–hipótese* é incorporado a um *prompt* fixo que solicita resposta binária (`0` ou `1`). Os parâmetros de geração são determinísticos (`temperature=0`, `do_sample=false`, `max_new_tokens=4`). A extração da predição usa a expressão regular `^[^a-z]*([01])`, capturando o primeiro dígito binário e tolerando caracteres de formatação iniciais. Respostas sem correspondência são contabilizadas como erros de conformidade.

A consistência intra-modelo foi avaliada em subconjunto fixo de 500 pares do ASSIN 2, reexecutados cinco vezes sob condições idênticas a fim de verificar a estabilidade das decisões e distinguir flutuações estocásticas de comportamento inferencial sistemático.

3.3. Modelos avaliados

A Tabela 1 apresenta os 11 modelos avaliados, organizados em três paradigmas. A execução foi realizada no ambiente Google Colab com GPUs NVIDIA L4 e A100, em precisão BF16 (FP8 para variantes *Instruct* do Ministral 3).

4. Resultados e Discussão

4.1. ASSIN 2: desempenho global

A Figura 1 apresenta a acurácia global dos modelos no ASSIN 2. A linha tracejada indica o piso heurístico (0,70) estabelecido por [da Silva et al. 2023]; a linha pontilhada, o teto supervisionado do BERTimbau-large (0,90), estado da arte na tarefa [Souza et al. 2020].

Tabela 1. Modelos avaliados, agrupados por paradigma.

Modelo	Tam.	Cobertura	Paradigma
Sabiá-7B	7B	PT-BR	Pré-treinamento contínuo
Bode-7B	7B	PT-BR	*Fine-tuning* (LoRA)
Gemma-3-Gaia-PT-BR-4B	4B	PT-BR	Pré-treino + *weight merging*
Ministral-3B-Instruct (FP8)	3B	Multilíngue	SFT + ODPO
Ministral-3B-Reasoning	3B	Multilíngue	GRPO em CoT
Ministral-8B-Instruct (FP8)	8B	Multilíngue	SFT + ODPO
Ministral-8B-Reasoning	8B	Multilíngue	GRPO em CoT
Qwen3-4B-Instruct	4B	Multilíngue	Híbrido *instruct*
Qwen3-4B-Thinking	4B	Multilíngue	Híbrido *thinking*
Gemma-2-2B-IT	2B	Multilíngue	Destilação de conhecimento
Gemma-2-9B-IT	9B	Multilíngue	Destilação de conhecimento

Modelos multilíngues recentes lideram: o Qwen3-4B-Instruct atinge 90,8%, equiparando-se ao teto supervisionado em regime *zero-shot*. O Gemma-2-9B e variantes *Reasoning* do Ministral também superam 87%. Modelos especializados em português exibem comportamento heterogêneo: o Gaia alcança 84,4%, mas Sabiá-7B e Bode-7B sofrem colapso de classe (acurácia global $\approx 50\%$, resultado sem direcionamento inferencial).

4.2. ASSIN 2: acurácia por classe

A Tabela 2 desagrega as acurácias por tipo de julgamento. Observa-se uma assimetria recorrente: a maioria dos modelos erra mais na classe *Não Acarretamento*, tendendo a aceitar hipóteses que não são formalmente acarretadas pela premissa. Sabiá-7B e Ministral-8B-Instruct colapsam para a classe negativa (0,0% e 0,7% em acarretamento); Bode-7B colapsa para a positiva (99,9% em acarretamento, 7,2% em não-acarretamento). A diferença de desempenho nas classes *Acarretamento* e *Não-Acarretamento* é significativa ($p < 0,0001$) para todos os modelos de acordo com o Teste Qui-Quadrado de Independência com Correção de Yates, com $\alpha = 0,05$.

4.3. ASSIN 2: acurácia por faixa de similaridade semântica

A Tabela 3 apresenta a acurácia segmentada pelas faixas de similaridade semântica (escala STS do ASSIN 2, de 1 a 5). Conforme definido por [Real et al. 2020], o valor 1 representa sentenças "totalmente diferentes" e o valor 5, "significado virtualmente idêntico". Nas faixas 1–2 (sentenças totalmente distintas), os melhores modelos atingem acurácia próxima de 100%, pois a decisão de não-acarretamento é trivial: a dissimilaridade total torna a relação dedutiva evidentemente falsa. O desafio intensifica-se nas faixas 3–4 e 4–5, onde a alta sobreposição lexical pode induzir modelos dependentes de similaridade a classificar incorretamente pares de alta semelhança como acarretamentos, mesmo quando a relação lógica não se sustenta.

O fenômeno mais crítico ocorre na faixa 5 (paráfrases): Sabiá-7B (14,3%) e Ministral-8B-Instruct (13,4%) colapsam justamente quando as sentenças são virtualmente idênticas, revelando viés sistemático pela classe negativa. O Bode-7B exibe o padrão in-

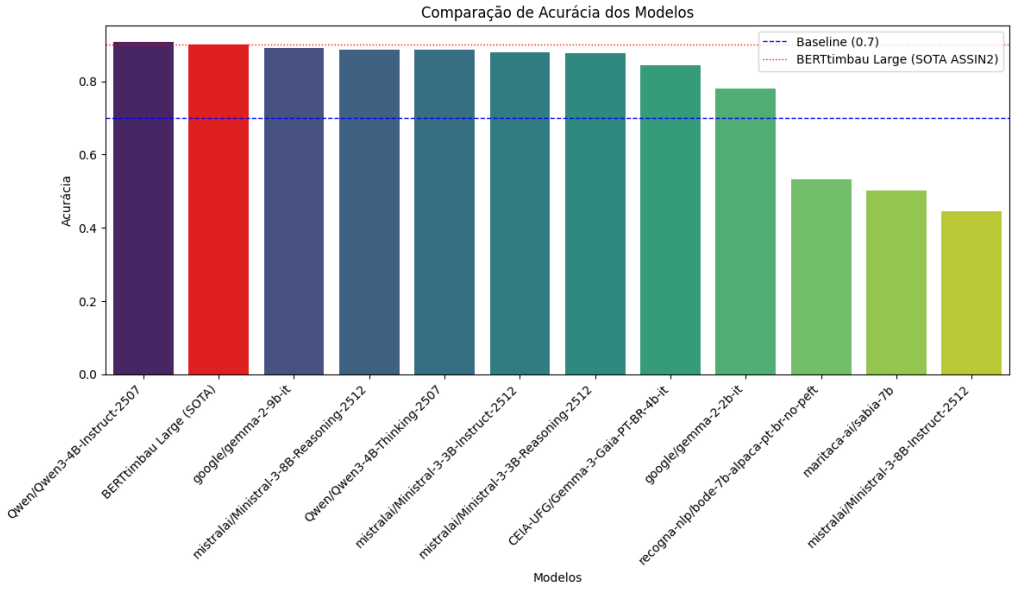


Figura 1. Acurácia global no ASSIN 2. Linha tracejada: piso heurístico (0,70); linha pontilhada: teto supervisionado BERTimbau-large (0,90).

verso: acerta bem na faixa 5 (86,5%) mas colapsa nas faixas 1–2 (10,1%), associando diretamente sobreposição lexical a acarretamento.

4.4. Dataset sintético: robustez sob controle semântico

A Tabela 4 apresenta os resultados no dataset sintético, construído exclusivamente com pares baseados em hiponímia. Nessa tabela, os valores de p referem-se à diferença de desempenho dos modelos nas classes *Acarretamento* e *Não Acarretamento*. Foram calculados aplicando-se o Teste Exato de Fisher (bicaudal), com $\alpha = 0{,}05$. O símbolo * indica significância estatística.

Qwen3-4B (*Instruct* e *Thinking*) e Gemma-2-9B mantêm acurácia $\geq$91,7%, demonstrando capacidade de generalização para a relação hiponímica sem dependência de regularidades do ASSIN 2. O Gaia também exibe robustez (90,0%), sugerindo que a especialização em português é mais eficaz quando combinada com uma base sólida de raciocínio. Em contraste, Sabiá, Bode e Ministral-8B-Instruct preservam o padrão de colapso, atingindo apenas 50%, acaso puro, em pares construídos para isolar a relação semântica formal.

4.5. Discussão

Padrões comportamentais e a distinção dedução/indução. A tipologia de Peirce oferece uma moldura interpretativa para os comportamentos observados. Modelos que colapsam para uma classe, Sabiá (sempre "não acarreta"), Bode (sempre "acarreta"), Ministral-8B-Instruct (sempre "não acarreta"), apresentam padrões de resposta incompatíveis com a verificação das condições de verdade que define o acarretamento: a decisão permanece constante independentemente do conteúdo proposicional do par. É importante ressaltar, contudo, que os experimentos medem exclusivamente comportamento observável em tarefas de classificação; inferir a partir dessas métricas a presença ou ausência de capacidades de raciocínio interno extrapola o que os dados permitem concluir diretamente. Esse

Tabela 2. Acurácia (%) por tipo de julgamento no ASSIN 2.

Modelo	Acarretamento	Não Acarret.	Global
Qwen3-4B-Instruct	94,7	86,8	**90,8**
Gemma-2-9B-IT	91,2	87,0	89,0
Ministral-8B-Reasoning	93,9	83,6	88,7
Qwen3-4B-Thinking	96,3	80,9	88,6
Ministral-3B-Instruct (FP8)	89,9	86,0	88,0
Ministral-3B-Reasoning	85,4	90,3	87,8
Gemma-3-Gaia-PT-BR-4B	95,9	72,9	84,4
Gemma-2-2B-IT	95,2	60,9	78,0
Bode-7B	99,9	7,2	53,4
Sabiá-7B	0,0	100,0	50,2
Ministral-8B-Instruct (FP8)	0,7	87,7	44,5

Tabela 3. Acurácia (%) por faixa de similaridade semântica (ASSIN 2).

Modelo	1–2	2–3	3–4	4–5	5	Global
Qwen3-4B-Instruct	98,5	93,3	92,4	88,3	91,3	90,8
Gemma-2-9B-IT	97,5	95,5	90,7	86,7	89,2	89,1
Ministral-8B-Reas.	94,9	91,7	89,3	85,7	89,9	88,7
Qwen3-4B-Thinking	99,5	96,2	91,2	83,5	89,9	88,6
Minist.-3B-Instr.(FP8)	95,5	95,7	90,8	85,5	87,8	88,0
Minist.-3B-Reas.	93,4	94,0	93,0	88,3	85,5	87,8
Gaia-PT-BR-4B	100,0	95,7	82,3	75,3	88,7	84,4
Gemma-2-2B-IT	98,5	82,6	69,3	65,9	86,3	78,0
Bode-7B	10,1	6,4	7,3	27,9	86,5	53,4
Sabiá-7B	100,0	99,8	97,6	78,9	14,3	50,2
Minist.-8B-Instr.(FP8)	95,5	94,3	86,5	67,3	13,4	44,5

comportamento reproduz as mesmas limitações das heurísticas simbólicas identificadas por [da Silva et al. 2023]: a distinção entre similaridade distribucional e preservação de condições de verdade permanece opaca para modelos sem mecanismos explícitos de raciocínio.

Raciocínio explícito e padrões de resposta mais robustos. Os melhores modelos, Qwen3, Gemma 2 e variantes *Reasoning* do Ministral, exibem padrões de resposta mais consistentes com os requisitos da tarefa: mantêm alto desempenho tanto nas faixas de baixa similaridade (rejeição correta de pares dissimilares) quanto na faixa 5 (aceitação correta de paráfrases), e preservam essa robustez nos dados sintéticos. Esses resultados sugerem que a *Chain-of-Thought* e a destilação de conhecimento produzem representações comportamentais mais alinhadas à distinção entre acarretamento e similaridade distribucional, embora não seja possível, a partir dos dados de classificação disponíveis, determinar se esse alinhamento resulta de raciocínio dedutivo genuíno ou de padrões aprendidos durante o treinamento que mimetizam esse comportamento.

Tabela 4. Acurácia (%) por tipo de julgamento no *dataset* sintético.

Modelo	Acarretamento	Não Acarret.	Global	p
Qwen3-4B-Thinking	90,0	100,0	**95,0**	0,2373
Qwen3-4B-Instruct	90,0	100,0	**95,0**	0,2373
Gemma-2-9B-IT	83,3	100,0	91,7	0,0522
Gemma-3-Gaia-PT-BR-4B	90,0	90,0	90,0	1,0000
Ministral-8B-Reasoning	73,3	100,0	86,7	0,0046 *
Ministral-3B-Instruct (FP8)	53,3	100,0	76,7	<0,0001 *
Ministral-3B-Reasoning	50,0	100,0	75,0	<0,0001 *
Gemma-2-2B-IT	96,7	40,0	68,3	<0,0001 *
Bode-7B	100,0	0,0	50,0	0 *
Ministral-8B-Instruct (FP8)	0,0	100,0	50,0	0 *
Sabiá-7B	0,0	100,0	50,0	0 *

A comparação entre Qwen3-4B-Thinking (88,6% no ASSIN 2 e 95,0% no sintético) e Qwen3-4B-Instruct (90,8% e 95,0%) indica que o *thinking mode*, embora marginalmente inferior no *benchmark*, é igualmente robusto em dados controlados. A superioridade do modo *instruct* no ASSIN 2 é estatisticamente significativa tanto no desempenho das classes *Acarretamento* e *Não-Acarretamento* isoladas quanto na média das duas ($p < 0,0001$ em Testes Qui-Quadrado de Homogeneidade, α = 0,05). Tal resultado parece sugerir uma exploração mais eficiente de padrões superficiais do *benchmark* por esse modelo. As diferenças se mostraram igualmente significativas nas comparações de mesmo tipo entre os modelos Gemma-2-9B-IT e Gemma-3-Gaia-PT-BR-4B. Entre eles, o ajuste fino em língua portuguesa do segundo modelo deve explicar os acertos na classe *Acarretamento*, negativamente compensada pelo mau desempenho na classe complementar. Considerando-se as duas classes, o modelo maior, sem ajuste fino específico para dados em português, acabou tendo resultados melhores globalmente.

A anomalia do Ministral-8B-Instruct. A disparidade entre Ministral-3B-Instruct (88,0%) e Ministral-8B-Instruct (44,5%) é contra-intuitiva perante as leis de escala. Com base nas informações do relatório técnico do Ministral 3 [Liu et al. 2026], uma hipótese explicativa é que o processo de alinhamento ODPO no modelo de 8B, ao otimizar a concisão e a segurança para assistentes de *chat*, possa ter suprimido padrões inferenciais mais elaborados, enquanto a destilação de *logits* específica aplicada ao modelo de 3B, combinada com a arquitetura de *tied embeddings*, teria mitigado esse efeito. É imprescindível, no entanto, ressaltar que essa interpretação é especulativa: não há experimentos de ablação ou avaliações controladas neste trabalho destinados a verificá-la empiricamente. Estudos futuros com variações sistemáticas dos parâmetros de alinhamento seriam necessários para confirmar ou refutar a hipótese.

Análise qualitativa de erros. A Tabela 5 ilustra como o colapso de classe se manifesta concretamente: Sabiá-7B nega o acarretamento mesmo quando a relação hiponímica é transparente (*lateral-esquerdo* $\subset$ *jogador*); Bode-7B afirma o acarretamento mesmo quando os predicados são de classes incompatíveis (*meia* $\not\to$ *árbitro*). O erro não de-

corre de dificuldades com o vocabulário, mas de viés decisório independente do conteúdo semântico do par.

Tabela 5. Predições corretas (✓) e incorretas (×) no dataset sintético.

Premissa	Hipótese	Acarret.	Sabiá	Bode
O lateral-esq. sofreu a falta	O jogador sofreu a falta	sim	(×)	(✓)
O meia organizou o meio-campo	O árbitro organizou o meio-campo	não	(✓)	(×)

***Zero-shot vs.* supervisionado.** O Qwen3-4B-Instruct (90,8%) iguala o BERTimbau-large ajustado supervisionadamente no ASSIN 2 (0,90), sem qualquer exemplo de treinamento. Isso demonstra que LLMs modernos compactos conseguem atingir o teto do estado da arte supervisionado, eliminando o custo de anotação manual. A diferença reside no custo de inferência: o BERTimbau-large opera com ∼335M parâmetros e pode rodar em hardware de entrada; o Qwen3-4B, mesmo quantizado, exige GPUs de pelo menos 8GB de VRAM. Para cenários de prototipagem rápida ou escassez de dados anotados, os LLMs *zero-shot* oferecem, portanto, uma alternativa viável ao ciclo custoso de anotação e treinamento supervisionado.

5. Conclusão

Este trabalho avaliou comparativamente 11 LLMs compactos em NLI em português brasileiro, em regime estritamente *zero-shot*, contrastando especialização linguística e raciocínio explícito. A introdução de um *dataset* sintético baseado em acarretamento por inclusão lexical (hiponímia/hiperonímia), permitiu separar competência inferencial genuína da exploração de regularidades superficiais do *benchmark*. Os resultados indicam que raciocínio explícito e destilação de conhecimento são mais decisivos do que adaptação monolíngue isolada para a tarefa. Do ponto de vista da semântica formal, modelos que exibem colapso de classe estão substituindo a verificação dedutiva de condições de verdade, a relação que define o acarretamento, pelo reconhecimento indutivo de similaridade distribucional. Acurácia elevada no ASSIN 2 pode, portanto, superestimar a competência inferencial quando não verificada em dados construídos sob restrição semântica estrita.

As principais limitações são: (i) o dataset sintético de tamanho reduzido e escopo temático único (60 pares, domínio do futebol, restrito à relação de hiponímia / hiperonímia), propõe-se mais como prova de conceito para avaliação controlada do que como avaliação de todos os tipos de acarretamento; (ii) os experimentos avaliam o comportamento observável em classificação binária e inferir a presença de raciocínio dedutivo interno a partir dessas métricas extrapola o que os dados permitem concluir diretamente e (iii) as interpretações causais sobre modelos específicos (notadamente o Ministral-8B-Instruct) são hipóteses baseadas em relatórios técnicos, não verificadas empiricamente neste estudo.

Como trabalhos futuros, vislumbra-se: a ampliação do dataset sintético para outras relações semânticas (negação, quantificação, pressuposição, predicação, operadores modais); a realização de estudos de ablação para investigar as causas do colapso em modelos específicos e a avaliação em regime *few-shot* para quantificar o ganho decorrente de exemplos explícitos.

Referências

Camilo-Junior, C. G., Oliveira, S. S. T., Pereira, L. A., Amadeus, M., Scotti, R., Fazzioni, D., Novais, A. M. A., and Jordão, S. A. A. (2025). GAIA: An open language model for Brazilian Portuguese. Hugging Face repository.

da Silva, F., Craveiro, G., da Silva, V., and Vanzin, V. (2023). Towards analysis on textual inference at ASSIN-2 dataset. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, pages 229–234. SBC.

Garcia, G. L., Paiola, P. H., Morelli, L. H., Candido, G., Cândido Júnior, A., Jodas, D. S., Afonso, L. C. S., Guilherme, I. R., Penteado, B. E., and Papa, J. P. (2024). Introducing Bode: A fine-tuned large language model for Portuguese prompt-based task. arXiv:2401.02909.

Gemma Team (2024). Gemma 2: Improving open language models at a practical size. arXiv:2408.00118.

Liu, A. H. et al. (2026). Ministral 3. arXiv:2601.08584.

Peirce, C. S. (1932). *Collected Papers of Charles Sanders Peirce, Volume II: Elements of Logic*. Harvard University Press, Cambridge, MA. Eds. Hartshorne, C.; Weiss, P.

Peirce, C. S. (1934). *Collected Papers of Charles Sanders Peirce, Volume V: Pragmatism and Pragmaticism*. Harvard University Press, Cambridge, MA. Eds. Hartshorne, C.; Weiss, P.

Pires, R., Abonizio, H., Almeida, T. S., and Nogueira, R. (2023). Sabiá: Portuguese large language models. In *Brazilian Conference on Intelligent Systems (BRACIS)*, pages 226–240. Springer.

Real, L., Fonseca, E., and Oliveira, H. G. (2020). The ASSIN 2 shared task: A quick overview. In *Proceedings of the International Conference on Computational Processing of the Portuguese Language (PROPOR)*, pages 406–412. Springer.

Real, L., Rodrigues, A., Vieira e Silva, A., Albiero, B., Thalenberg, B., Guide, B., Silva, C., de Oliveira Lima, G., Câmara, I. C. S., Stanojević, M., Souza, R., and de Paiva, V. (2018). SICK-BR: A Portuguese corpus for inference. In *Proceedings of the International Conference on Computational Processing of the Portuguese Language (PROPOR)*, pages 303–312. Springer.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: Pretrained BERT models for Brazilian Portuguese. In *Brazilian Conference on Intelligent Systems (BRACIS)*, pages 403–417. Springer.

Yang, A. et al. (2025). Qwen3 technical report. arXiv:2505.09388.

Quando o "..." importa: análise comparada de modalidades de transcrição de hesitações em entrevistas do Corpus Roda Viva

Valeria Vieira dos Santos

Programa de Pós-Graduação em Linguística (PPGL)
Universidade Federal de São Carlos (UFSCar)

valeriavsantos93@gmail.com

Abstract. *This paper investigates how hesitations — pauses, lengthenings, truncations, repetitions, and lexical fillers — are represented across different transcription modalities in Brazilian Portuguese journalistic interviews. Drawing on a pilot corpus of 95 speaker turns from the Roda Viva program (TV Cultura), we compare a faithful transcription (preserving hesitation markers) with a sanitized version and a conversational-protocol transcription (Jefferson; GAT2). Results show that each modality systematically suppresses, normalizes, or preserves hesitation markers in distinct ways, revealing the need for an explicit annotation protocol. We propose an initial functional typology of hesitations in five formal categories and discuss implications for oral corpus construction and NLP annotation tasks.*

Resumo. *Este trabalho investiga como as hesitações — pausas, prolongamentos, truncamentos, repetições e preenchimentos lexicais — são representadas em diferentes modalidades de transcrição de entrevistas jornalísticas do português brasileiro. A partir de um corpus piloto de 95 turnos de fala do programa Roda Viva (TV Cultura), comparamos uma transcrição fiel (com marcas hesitativas) com uma versão higienizada e com uma transcrição em protocolos conversacionais (Jefferson; GAT2). Os resultados mostram que cada modalidade trata sistematicamente as marcas hesitativas de forma distinta, revelando a necessidade de um protocolo de anotação explícito. Propomos uma tipologia funcional inicial em cinco categorias formais e discutimos implicações para a construção de corpora orais e tarefas de anotação em PLN.*

1. Introdução

A transcrição de fala espontânea é uma etapa fundamental na construção de corpora orais computacionalmente utilizáveis. Diferentes práticas de transcrição, como jornalísticas, automáticas e conversacionais, são orientadas por objetivos distintos e produzem

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 522–530
October 19-22, 2026 ©2026 Association for Computational Linguistics

representações diversas de um mesmo evento de fala, com consequências diretas sobre a qualidade e a completude dos recursos gerados (Vale et al., 2024). Entre os fenômenos mais afetados por essas diferenças estão as *hesitações*: pausas, prolongamentos, truncamentos, preenchimentos lexicais (como *éé́é*, *ahn*, *então*) e repetições. Elementos recorrentes da fala espontânea que representam aproximadamente 6% dos tokens em interações orais (Fox Tree, 1995; Shriberg, 2001).

A transcrição **jornalística** tem como objetivo a legibilidade pública: remove marcas interpretadas como ruído. A transcrição **automática**, gerada por sistemas como YouTube/Google, produz texto fluente em larga escala, mas com baixa sensibilidade a fenômenos disfluentes. Já as transcrições **conversacionais**, com protocolos como Jefferson (2004) e GAT2 (Selting et al., 2009), registram com precisão a materialidade da fala para fins analíticos. O resultado é que as hesitações são apagadas em algumas modalidades, normalizadas em outras, e detalhadamente preservadas apenas nos protocolos especializados, criando heterogeneidade que compromete a consistência do corpus para uso em PLN.

Este trabalho parte de um corpus piloto de 95 turnos de fala coletados de uma entrevista do programa *Roda Viva* (TV Cultura), integrante do Corpus Roda Viva (Vale et al., 2024), e tem como objetivo: (i) comparar sistematicamente o tratamento das hesitações em duas modalidades de transcrição, fiel e higienizada; (ii) contrastar com exemplos em protocolos conversacionais; e (iii) propor uma tipologia funcional inicial das hesitações com base nas diferenças observadas. O objetivo final é contribuir com fundamentos linguísticos para a construção de um protocolo de anotação explícito e reutilizável para o Corpus Roda Viva.

2. Hesitação como fenômeno linguístico

As hesitações são fenômenos formalmente descritíveis: cada tipo possui propriedades distribucionais, prosódicas e funcionais que permitem categorização sistemática (Shriberg, 2001; Corley & Stewart, 2008). Não se trata de ruído ou falha de desempenho, mas de recursos da fala que cumprem funções específicas na organização do discurso oral: indicam planejamento linguístico em curso, sinalizam dificuldades lexicais ou estruturais, e participam da gestão de turnos em interações assimétricas como entrevistas jornalísticas.

Com base na literatura (Shriberg, 2001; Corley & Stewart, 2008; Zayats & Ostendorf, 2019, Marcuschi, 2003) e nos dados do corpus piloto, identificamos cinco categorias formais de hesitação:

Pausas, que são os intervalos de silêncio dentro do turno de fala, sem transferência de turno ao interlocutor. Podem ser breves (.) ou longas (..) e ocorrem em posições de planejamento discursivo, fronteiras sintáticas ou momentos de busca lexical (Shriberg,

2001). Em protocolos conversacionais como Jefferson (2004), sua duração é estimada; nas transcrições jornalísticas e automáticas, são sistematicamente apagadas.

Preenchimentos lexicais (fillers), vocalizações como éééé, ahn, e usos hesitantes de então e né funcionam como sinais de que o falante mantém o turno enquanto planeja o enunciado (Marcuschi, 2003; Corley & Stewart, 2008). São os marcadores mais frequentemente apagados em transcrições jornalísticas.

Truncamentos, representados por interrupções abruptas de palavras antes de seu término, geralmente seguidas de reformulação (ex.: Mina- Minas, dem- práticas). Indicam replanejamento imediato e são representados com hífen nos protocolos conversacionais, mas frequentemente omitidos ou corrigidos nas demais modalidades.

Prolongamentos, sendo as extensões anormais de sílabas, representadas por :: (ex.: Então::). Indicam hesitação prosódica e funcionam como estratégia para manter o turno. São raramente preservados em transcrições que não seguem protocolos especializados.

E Repetições, reiterações de sílabas, palavras ou sintagmas em contexto de replanejamento discursivo (ex.: do do, para... para...). Marcuschi (2003) as distingue de repetições retóricas intencionais pela co-ocorrência com outros marcadores hesitativos e pela ausência de função enfática.

A Tabela 1 sintetiza as cinco categorias com seus marcadores e funções:

Tabela 1. Tipologia funcional inicial das hesitações — categorias e marcadores

Categoria	Marcador (ex.)	Função / descrição
Pausa	(.) (..) (...)	Intervalo de silêncio dentro do turno; pode ser breve (.) ou longa (..)
Preenchimento lexical	éééé, ahn, éé, então	Vocalizações que preenchem pausas e sinalizam continuidade de turno
Truncamento	palavra- (ex.: Mina-)	Interrupção abrupta de palavra, seguida de reformulação ou nova tentativa
Prolongamento	:: (ex.: Então::)	Extensão anormal de sílaba; indica hesitação prosódica e busca lexical
Repetição	do do, na na, para... para...	Reiteração de sílaba, palavra ou sintagma; marca replanejamento

3. Modalidades de transcrição e seus protocolos

Cada modalidade de transcrição é orientada por um objetivo diferente, o que determina diretamente o tratamento das hesitações. A transcrição **jornalística** prioriza a inteligibilidade do conteúdo: edita a fala para parecer fluente, apagando sistematicamente

preenchimentos, pausas e truncamentos. A transcrição **automática**, gerada por sistemas de reconhecimento automático de fala como os do YouTube/Google, otimiza velocidade e cobertura, produzindo texto normalmente sem marcas de oralidade. Essas limitações já foram documentadas para o português brasileiro por Lima & Campelo (2024) e Souza et al. (2025).

Já os protocolos conversacionais têm propósito analítico: o **sistema Jefferson** (2004) registra a estrutura interacional da fala (turnos, sobreposições, pausas cronometradas), e o **GAT2** (Selting et al., 2009) acrescenta detalhamento prosódico (entonação, ênfase, intensidade). Juntos, esses protocolos oferecem a representação mais fiel da hesitação como fenômeno da fala espontânea, mas são trabalhosos e raramente aplicados a larga escala.

A Tabela 2 sintetiza as diferenças entre as três modalidades em relação ao tratamento das hesitações.

Tabela 2. Tratamento das hesitações por modalidade de transcrição

Categoria	Jornalística	Automática	Conversacional
Pausa	Apagada	Apagada	Marcada e cronometrada
Preenchimento lexical	Apagado	Parcialmente mantido	Preservado e transcrito
Truncamento	Apagado ou corrigido	Parcialmente mantido	Marcado com hífen
Prolongamento	Apagado	Apagado	Marcado com ::
Repetição	Apagada	Apagada	Preservada

4. Corpus e metodologia

O corpus desta pesquisa é o Corpus Roda Viva, coordenado pelo Prof. Dr. Oto Araújo Vale (UFSCar), que organiza 713 entrevistas do programa *Roda Viva* (TV Cultura) com transcrições jornalísticas, arquivos audiovisuais e versões em Universal Dependencies (Vale et al., 2024). Para este trabalho, utilizamos um corpus piloto composto por **95 turnos de fala** extraídos de uma entrevista completa, com dois locutores principais e mais cinco participantes, totalizando sete locutores distintos. Por se tratar de uma única entrevista, os resultados aqui apresentados têm caráter exploratório; a ampliação da amostra, com seleção intencional de entrevistas por período, área temática e perfil de entrevistados, é etapa prevista da pesquisa em andamento.

O corpus piloto foi organizado em três colunas: (i) **FALA_FIEL**: transcrição com marcadores de hesitação preservados, seguindo convenções próximas ao sistema Jefferson; (ii) **FALA_HIGIENIZADA**: versão sem marcas hesitativas, equivalente a uma transcrição jornalística; e (iii) **NOTA_HUMANA**: score de 0 a 100 indicando o grau de semelhança entre as duas versões, atribuído manualmente (100 = idênticas; 0 = máxima divergência). Adicionalmente, para fins de exemplificação qualitativa, utilizamos transcrições nos protocolos Jefferson e GAT2 de uma entrevista publicada anteriormente (Vale et al., 2024).

A NOTA_HUMANA foi atribuída manualmente como medida de similaridade percebida entre a FALA_FIEL e a FALA_HIGIENIZADA de cada turno: valores próximos de 100 indicam que as duas versões são praticamente idênticas, ou seja, o turno não continha marcas hesitativas relevantes; valores próximos de 0 indicam máxima divergência, com apagamento extenso de marcadores. A nota foi atribuída por uma única anotadora, linguista, mestre em Linguística, com experiência em transcrição de fala e Análise da Conversação. Optou-se pela escala contínua de 0 a 100, em vez de escalas ordinais do tipo Likert, por sua maior granularidade e pela compatibilidade direta com métricas computacionais de similaridade normalizadas (0–1), que serão empregadas na etapa subsequente da pesquisa. Trata-se, ainda assim, de uma métrica exploratória, baseada em julgamento humano, que permite estratificar o corpus piloto por densidade hesitativa sem depender de contagem automática e que será substituída por esquemas de anotação validados por Acordo Inter-Anotadores (IAA).

As transcrições conversacionais foram produzidas com auxílio do software ELAN (Wittenburg et al., 2006), que permite o alinhamento entre anotação textual e os arquivos de áudio e vídeo das entrevistas, viabilizando a verificação direta das marcas hesitativas no sinal acústico original.

A análise quantitativa consistiu em: (i) contagem de ocorrências por categoria de hesitação na FALA_FIEL; (ii) verificação de supressão na FALA_HIGIENIZADA; e (iii) análise da distribuição do score NOTA_HUMANA para medir o impacto do apagamento.

5. Análise comparada

O corpus piloto contém **903 marcadores de hesitação** na FALA_FIEL, distribuídos conforme a Tabela 3. Todos são sistematicamente apagados ou normalizados na FALA_HIGIENIZADA.

Tabela 3. Frequência de marcadores hesitativos no corpus piloto (95 turnos)

Categoria	Ocorrências	%
Truncamento (palavra-)	289	32,0
Pausa breve (.)	264	29,2

Preenchimento (éé́é, ahn)	247	27,4
Prolongamento (::)	75	8,3
Pausa longa (..)	28	3,1
Total	**903**	**100**

A distribuição do score NOTA_HUMANA revela o impacto do apagamento: 26 turnos (27,4%) receberam nota 0, ou seja, as duas versões foram consideradas completamente divergentes, indicando que praticamente todas as marcas hesitativas foram removidas. Outros 14 turnos (14,7%) receberam nota 100, correspondendo a falas sem hesitação registrada. Os demais 55 turnos (57,9%) apresentaram divergência parcial, com apagamento seletivo de marcadores específicos.

Esses resultados têm implicação direta para a construção do corpus: turnos com nota 0 representam os casos de maior perda informacional na passagem da modalidade fiel para a jornalística, e são também os mais relevantes para o treinamento de modelos de detecção automática, pois concentram a maior densidade de marcadores por turno. Já os turnos com nota 100 correspondem majoritariamente a falas curtas e formulaicas (cumprimentos, confirmações, agradecimentos), nas quais a ausência de hesitação é esperada pelo contexto interacional. A faixa intermediária (notas entre 40 e 90, com 40 turnos) revela os casos de apagamento parcial e inconsistente, onde a escolha de quais marcadores preservar ou remover não segue critério explícito, evidenciando precisamente a lacuna que um protocolo de anotação formal viria preencher.

O Exemplo 1 ilustra o apagamento seletivo de marcadores na transição da FALA_FIEL para a FALA_HIGIENIZADA (ID 16, locutor jornalista, nota 65):

> **Exemplo 1 — Fala fiel:** *boa noite Luísa (.) boa noite a todos da roda **éé́é** noite queria aproveitar esse gancho [...] **éé́é** você fala tutela militar né (.) e no- nas entrevistas...*
>
> **Fala higienizada:** *boa noite Luísa boa noite a todos da roda noite queria aproveitar esse gancho [...] você fala tutela militar né e no- nas entrevistas...*

Nota-se que os dois preenchimentos *éé́é* e a pausa breve *(.)* foram completamente removidos. O truncamento *no-* foi preservado na versão higienizada neste caso, o que ilustra a **inconsistência** no tratamento de categorias distintas dentro de uma mesma versão, evidenciando a ausência de um protocolo explícito.

O Exemplo 2 demonstra como o mesmo trecho é representado nas três modalidades (a partir de entrevista disponível em Vale et al., 2024):

> **Exemplo 2 — Três modalidades do mesmo trecho de fala:**

[Automática] *"Eu quero dizer, talvez a gente... três eixos grandes. Um: nós temos que, as forças armadas têm que apresentar o país, e o país tem que discutir qual é o plano de defesa."*

[Jefferson] *"Eu. Eu quero dizer. . . talvez a gente – isso – 3, 3.. eixos grandes, um (.). Nós temos que /as forças armadas têm que apresentar o país.. e o país tem que discutir.. qual é o plano de defesa?"*

[GAT2] *"Eu (.) eu quero dizer. . . talvez a gente – isso (.) três, três.. EIXOS grandes (.) um (.) nós temos que /as forças armadas têm que apresentar o país (.) e o país tem que disCUTIR (.) qual é o plano de DE FESA?↗"*

A versão automática apaga as repetições (*Eu. Eu*), as pausas e os truncamentos. A versão Jefferson registra a estrutura interacional completa. A versão GAT2 acrescenta ênfase prosódica e direção de entonação. Cada modalidade disponibiliza um recorte diferente da mesma realidade linguística, e a escolha de modalidade determina quais fenômenos existem ou não no corpus.

6. Para uma tipologia funcional

Os dados apresentados sustentam a necessidade de uma tipologia funcional das hesitações que seja: (i) **formalmente descritiva**, com critérios explícitos para cada categoria, independentes de julgamentos subjetivos; (ii) **multimodalmente compatível**, mapeável entre as três modalidades de transcrição; e (iii) **computacionalmente operacionalizável**, exportável em schema legível por máquina para uso em ferramentas de anotação e treino de modelos de linguagem.

A Tabela 1, apresentada na Seção 2, constitui uma proposta inicial nessa direção. Sua validação metodológica , via Teste de Acordo Inter-Anotadores (IAA) com Kappa de Cohen e Alpha de Krippendorff, está prevista como etapa subsequente da pesquisa, assim como sua implementação em formato ELAN/Universal Dependencies para integração ao Corpus Roda Viva.

Vale destacar também a relevância da tipologia para tarefas de PLN que vão além da anotação de corpora. A detecção automática de hesitações tem aplicação direta em sistemas de legendagem automática, que hoje produzem legendas "limpas" sem marcas de oralidade, em ferramentas de análise de entrevistas jornalísticas em larga escala, e no desenvolvimento de modelos de linguagem mais sensíveis à fala espontânea. Trabalhos recentes demonstram que LLMs aplicados a transcrições em português brasileiro cometem erros sistemáticos justamente nas categorias que as transcrições automáticas apagam: truncamentos e preenchimentos lexicais (Lima & Campelo, 2024). Uma tipologia formalmente descrita e um corpus de referência anotado são, portanto, pré-requisitos para que esses modelos possam ser treinados e avaliados de forma consistente, e não apenas testados em dados que já não contêm o fenômeno que se quer detectar.

Do ponto de vista computacional, os dados quantitativos do corpus piloto (Tabela 3) já fornecem evidências sobre quais categorias são mais frequentes e, portanto, mais relevantes para treino de modelos de detecção automática. O alto índice de truncamentos (32,0%) e preenchimentos lexicais (27,4%) sugere que essas categorias constituem prioridade para a anotação de referência a ser desenvolvida.

7. Considerações finais

Este trabalho demonstrou, a partir de um corpus piloto de 95 turnos de fala do Corpus Roda Viva, que as modalidades de transcrição jornalística, automática e conversacional tratam as hesitações de forma sistematicamente distinta — apagando, normalizando ou preservando marcadores de acordo com seus objetivos. Em 27,4% dos turnos analisados, o apagamento foi total; em 57,9%, foi parcial e inconsistente, sem critério explícito.

A proposta de tipologia funcional em cinco categorias (pausa, preenchimento lexical, truncamento, prolongamento e repetição) oferece um ponto de partida para um protocolo de anotação explícito, replicável e compatível com múltiplas modalidades de transcrição. A consolidação dessa tipologia, incluindo validação por IAA e implementação em formato legível por máquina, está prevista como contribuição metodológica central da pesquisa em andamento, cujos produtos serão integrados ao repositório público do Corpus Roda Viva.

Os dados e exemplos apresentados reforçam que a escolha da modalidade de transcrição não é neutra: ela determina quais fenômenos linguísticos existem no corpus. Tornar essa escolha explícita e os critérios de anotação públicos é condição necessária para que corpora orais sejam utilizáveis de forma consistente em pesquisa linguística e em PLN.

Cabe destacar que o corpus piloto quantitativo é derivado de uma única entrevista, o que limita a generalização dos resultados — ainda que a comparação qualitativa entre modalidades (Exemplo 2) recorra a uma segunda entrevista do corpus. A ampliação da amostra, com seleção intencional por período, área temática e perfil de entrevistados, é etapa prevista da pesquisa.

Referências

Corley, M. and Stewart, O. W. (2008) "Hesitation disfluencies in spontaneous speech: the meaning of um." Language and Linguistics Compass, v. 2, n. 4, p. 589–602.

Fox Tree, J. E. (1995) "The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech." Journal of Memory and Language, v. 34, n. 6, p. 709–738.

Jefferson, G. (2004) "Glossary of transcript symbols with an introduction." In: Lerner, G. H. (Ed.). Conversation Analysis: Studies from the first generation. Amsterdam: John Benjamins. p. 13–31.

Lima, P. L. S. de and Campelo, C. E. C. (2024) "Disfluency Detection and Removal in Speech Transcriptions via Large Language Models." In: Anais do XIV STIL. Porto Alegre: SBC. Disponível em: https://aclanthology.org/2024.stil-1.16.

Marcuschi, L. A. (2003) Análise da conversação. 5. ed. São Paulo: Ática.

Selting, M. et al. (2009) "Gesprächsanalytisches Transkriptionssystem 2 (GAT 2)." Gesprächsforschung, v. 10, p. 353–402.

Shriberg, E. (2001) "To 'errrr' is human: ecology and acoustics of speech disfluencies." Journal of the International Phonetic Association, v. 31, n. 1, p. 153–169.

Souza, K. D. F. de, Pereira, D. E., Campelo, C. E. C. and Vasconcelos, L. L. (2025) "DEBISS: a Corpus of Individual, Semi-structured and Spoken Debates." arXiv, 2603.05459. Disponível em: https://arxiv.org/abs/2603.05459.

Vale, O. A. et al. (2024) "Roda Viva Corpus: An Audiovisual Brazilian Portuguese Interview Corpus." In: Proceedings of PROPOR 2024. p. 204–210. Disponível em: https://aclanthology.org/2024.propor-2.22.

Wittenburg, P. et al. (2006) "ELAN: a professional framework for multimodality research." In: Proceedings of LREC 2006. Genova: ELRA. p. 1556–1559.

Zayats, V. and Ostendorf, M. (2019) "Giving attention to the unexpected: using prosody innovations in disfluency detection." In: Proceedings of NAACL 2019. Minneapolis: ACL. p. 2162–2172.

Da sintaxe ao discurso: Evidências introdutórias da correlação entre *Universal Dependencies* e *Rhetorical Structure Theory*

Jackson Wilke da Cruz Souza[1], Paula Christina Figueira Cardoso[2]

[1]Programa de Pós-Graduação em Língua e Cultura
Universidade Federal da Bahia - Salvador/BA

[2]Instituto de ciências exatas e naturais - Faculdade de Computação
Universidade Federal do Pará – Belém/PA

{jacksoncruz@ufba.br; pcardoso@ufpa.br}

***Abstract**: The automatic identification of discourse relations constitutes a central challenge in Natural Language Processing, particularly for Brazilian Portuguese, due to the scarcity of annotated resources and the variability of discourse signaling mechanisms. In this preliminary study, we investigate the correlation between discourse relations from Rhetorical Structure Theory (RST) and syntactic dependency relations within the Universal Dependencies (UD) framework in journalistic texts. We analyze five RST relations that are frequent in this genre, combining quantitative and qualitative analyses. The results indicate that each discourse relation exhibits distinct syntactic-discursive characteristics.*

***Resumo:** A identificação automática de relações discursivas constitui um desafio central no Processamento de Línguas Naturais, especialmente para o português brasileiro, devido à escassez de recursos anotados e à variabilidade dos mecanismos de sinalização discursiva. Neste trabalho introdutório investigamos a correlação entre relações discursivas da Rhetorical Structure Theory (RST) e relações de dependência sintática do modelo Universal Dependencies (UD) em textos jornalísticos. Analisamos cinco relações RST frequentes no gênero, combinando análises quantitativas e qualitativas. Os resultados indicam que cada relação discursiva apresenta características sintático-discursivas distintas.*

1. Introdução

O nível discursivo de análise linguística impõe uma série de desafios a pesquisadores da área de Processamento de Línguas Naturais (PLN), já que muitos fenômenos exigem mais poder de processamento e/ou descrição visando a implementação computacional. Além disso, há escassez de recursos linguístico-computacionais robustos capazes de lidar adequadamente com esse nível de análise, sobretudo no contexto do português do Brasil (PB).

A *Rhetorical Structure Theory* (RST) [Mann e Thompson, 1988] tem sido bastante utilizada em tarefas de PLN, como sumarização automática, detecção de notícias falsas e avaliação automática de redações [Souza, Cardoso e Rodrigues, 2023], bem como auxiliando

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 531–539
October 19-22, 2026 ©2026 Association for Computational Linguistics

propriamente na análise discursiva (Liu *et al.*, 2021). Tais pesquisas, de maneira geral, utilizam marcadores discursivos (MDs), tipicamente conjunções e preposições, para rotular possíveis relações entre as *Elementary Discourse Units* (EDUs), tidas como unidades mínimas de análise.

Em consonância a trabalhos de destaque internacional, Rodrigues, Souza e Cardoso (2023) e Cardoso *et al.* (2024) identificaram que relações retóricas em português também podem ocorrer entre EDUs sem a necessidade de marcadores discursivos tradicionalmente descritos. Esses estudos adotaram a noção de Sinalizadores Discursivos (SDs), em que diferentes categorias de pistas linguístico-estruturais podem contribuir para a identificação de relações discursivas. Neste estudo, entretanto, restringimo-nos a sinalizadores de natureza sintática, operacionalizados por meio das relações de dependência propostas pela *Universal Dependencies* (UD) [Nivre, 2015], por constituírem o foco da investigação aqui desenvolvida. Essa delimitação permite investigar, de forma mais sistemática, possíveis regularidades na interface entre sintaxe e discurso.

Na UD, as relações sintáticas são definidas a partir da dependência entre elementos que compõem a sentença, permitindo a extração de informações gramaticais no nível morfossintático. No cenário do PLN do Português, o modelo UD vem sendo aplicado em diferentes gêneros textuais, além de estar agregado a uma gama de recursos (como *corpora* e léxicos) e ferramentas (como analisadores e modelos pré-treinados de anotação automática) [Di-Felippo *et al.*, 2021].

Partindo do princípio de que há uma correlação entre Discurso e Sintaxe, nosso objetivo neste trabalho é investigar possíveis associações entre padrões sintáticos da UD e relações discursivas da RST, buscando identificar regularidades recorrentes. Para tanto, realizamos dois tipos de análises. Na primeira, de base qualitativa, investigamos como e quando relações dos dois modelos estão correlacionadas; na segunda, de base quantitativa, buscando verificar a recorrência de relações UD e como poderia resultar em padrões associáveis a relações RST. Este trabalho está organizado em 4 seções, além desta Introdução. Na Seção 2, apresentamos os trabalhos relacionados ao tema desta pesquisa. Na Seção 3 delimitamos os procedimentos metodológicos aqui empreendidos. Na Seção 4 demonstramos os resultados. Por fim, na Seção 5, tecemos nossas considerações finais.

2. Trabalhos relacionados

Conforme Cardoso, Souza e Rodrigues (2024), o nível discursivo pode ser compreendido como uma confluência de outros níveis de análise linguística, o que implica a ocorrência de fenômenos situados em zonas de fronteira entre esses níveis. No âmbito do PLN, a relação entre sintaxe e discurso tem sido amplamente investigada, destacando-se estudos que articulam esses domínios a partir de modelos teóricos como a RST e a UD.

Nesse contexto, Zeldes (2024) desenvolveu a ferramenta *rst2dep*, voltada à conversão entre representações discursivas da RST e estruturas sintáticas da UD. Um dos principais desafios enfrentados nessa proposta diz respeito ao pareamento das unidades de análise, uma vez que a RST opera sobre EDUs, enquanto a UD se baseia em tokens. Para lidar com essa assimetria, a ferramenta parte do pressuposto de que o discurso também apresenta organização estrutural, propondo estratégias para a identificação de EDUs nucleares, inclusive em construções multinucleares. Para tanto, o autor parte do princípio de que uma das soluções relevantes seja a transposição entre os modelos,

considerando que a UD constrói sua representação arbórea a partir de um nó raiz (root), do qual derivam todas as relações sintáticas.

Outra linha de investigação profícua em PLN que articula os níveis discursivo e sintático concentra-se no estudo de conectivos e MDs. A centralidade desses elementos decorre de seu papel como sinalizadores de relações discursivas, aliado à sua ampla diversidade morfossintática. Em diversos casos, tais sinalizadores exploram relações sintáticas, sobretudo intersentenciais, para evidenciar relações discursivas. À título de exemplo tem-se o que ocorre com o conector *mas*, que, ao mesmo tempo pode expressar contraposição sintática, mas também pode sinalizar contraste discursivo entre EDUs.

De acordo com Zufferey e Degand (2024), abordagens fundamentadas em modelos como a RST e representações de dependência evidenciam a presença de conectivos, embora estes não constituam condição necessária para a identificação de relações discursivas. Muitas relações podem ser inferidas implicitamente a partir de informações estruturais, semânticas e pragmáticas, o que reforça a distinção entre marcação explícita e coerência discursiva subjacente [Das e Taboada, 2014; Rodrigues, Souza e Cardoso, 2023].

Do ponto de vista computacional, essas considerações têm implicações diretas para o desenvolvimento de *parsers* discursivos e sistemas automáticos de análise do discurso. A ambiguidade funcional de conectivos e sua dependência do contexto indicam a necessidade de modelos integrados, capazes de articular informações sintáticas e discursivas de forma consistente. Assim, o estudo de conectivos constitui um eixo central para a compreensão e a modelagem da interface entre sintaxe e discurso em PLN.

A escolha do modelo sintático adotado neste trabalho justifica-se pelo fato de a UD oferecer um conjunto unificado de padrões para a representação sintática em diferentes línguas, o que possibilita comparações interlinguísticas [Freitas, 2025]. Ademais, para o português, há ferramentas treinadas e amplamente disponíveis para a anotação e análise automática de estruturas de dependência, tanto para textos formais [Lopes e Pardo, 2024] quanto para *corpora* multigênero [Di-Felippo *et al.,* 2024].

Mais recentemente, *Large Language Models* (LLMs) também passaram a ser empregados em tarefas de segmentação e *parsing* discursivo, alcançando resultados robustos na identificação automática de relações retóricas por meio de estratégias de *fine-tuning* e *prompting* [Maekawa *et al.*, 2024; Thompson *et al.*, 2024]. Apesar de essas abordagens representem avanços significativos para a área de PLN, a análise discursiva permanece pouco interpretável, o que dificulta a compreensão de quais propriedades linguístico-estruturais efetivamente contribuem para a identificação das relações discursivas. Assim, ao passo que trabalhos recentes se concentram no aprimoramento do desempenho de *parsers* discursivos, este estudo busca compreender, ainda que sucintamente, quais possíveis regularidades sintáticas podem explicar a ocorrência de determinadas relações retóricas, oferecendo evidências linguísticas que podem, futuramente, subsidiar sistemas automáticos de análise discursiva.

3. Metodologia

Para o desenvolvimento deste trabalho, separamos uma amostra de sentenças anotadas previamente com o modelo RST provenientes do *corpus* CSTNews [Cardoso *et al.*, 2011], composto por textos jornalísticos em língua portuguesa. Por conta do gênero textual, construímos a amostra em função de relações que são mais características, a saber, *Sequence*, *Elaboration*, *Purpose*, *Attribution* e *Circumstance.*

O conjunto de sentenças foi submetido à ferramenta PortParserv2 [Lopes e Pardo, 2024], que realiza anotação UD automaticamente. A escolha por essa ferramenta se deu por conta de ela ter sido pré-treinada com textos jornalísticos. Essa dupla anotação (discursiva, via RST, e sintática, via UD) permite a investigação sistemática da interface entre sintaxe e discurso.

A partir das anotações UD, foram extraídos padrões sintáticos, definidos como sequências ordenadas de rótulos de dependência, respeitando a ordem dos tokens na sentença. Esses padrões representam abstrações da estrutura sintática e permitem a comparação entre sentenças de diferentes relações discursivas sem recorrer a informações lexicais específicas.

Para cada subconjunto correspondente a uma relação RST, foram computadas as frequências absolutas e relativas dos padrões UD observados. Em seguida, os padrões mais frequentes em cada relação foram selecionados, com o objetivo de identificar regularidades sintáticas potencialmente associadas a funções discursivas específicas.

Ainda, a abstração dos padrões consistiu na eliminação de modificadores opcionais da organização linear de constituintes e distinções exclusivamente morfológicas (como `amod` e `advmod`, quando não alteravam a configuração principal). Essa decisão baseou-se na ideia de preservar apenas as relações de dependência consideradas estruturantes da configuração sintática. O objetivo foi reduzir a esparsidade dos dados sem eliminar informações relevantes para caracterizar cada relação discursiva.

4. Resultados e discussão

Na Tabela 1, apresentamos a síntese dos resultados obtidos a partir da análise realizada. Nela reunimos os macro-padrões recorrentes nos trechos correspondentes para cada relação RST, organizados em termos de porcentagem (P). Os padrões de dependência foram extraídos a partir das anotações UD completas, agrupando-se sequências estruturalmente equivalentes por meio da eliminação de modificadores opcionais, variações de ordem linear e distinções morfológicas não centrais. Essa abstração permitiu reduzir a esparsidade dos dados e evidenciar regularidades sintático-discursivas associadas às relações RST analisadas.

Os números absolutos de exemplos para cada relação são: *Purpose* (32), *Attribution* (30), *Sequence* (25), *Elaboration* (20) e *Circumstance* (30).

Destaca-se que os padrões sintáticos apresentados na Tabela 1 não correspondem necessariamente a sequências fixas de palavras, mas sim a estruturas sintáticas representadas pelas relações de dependência da UD. Nesse caso, um mesmo padrão pode ser realizado por diferentes construções frásticas, desde que preserve a organização estrutural de seus constituintes. Nos trechos exemplificados, extraídos do *corpus* CSTNews, observa-se que relações como `case + det + obl (+nmod)` ou `mark + advcl` são instanciadas por diferentes realizações linguísticas, evidenciando que o foco da representação está sobre a função sintática dos constituintes e suas dependências, e não sobre combinações léxicas.

A relação *Sequence*, na RST, ocorre quando eventos ou estados são organizados em ordem temporal, procedural ou narrativamente, sem hierarquia entre as EDUs, sendo, portanto, uma relação multinuclear. Em textos jornalísticos, esse tipo de organização discursiva manifesta-se sobretudo por meio de coordenação de predicações, parataxe, marcas temporais e aspectuais e paralelismo sintático.

Relação RST	Padrão UD abstrato	P (%)	EXEMPLO
Purpose	`mark + advcl`	61	“quando Abadia se entregou à polícia colombiana”
	`case + det + obl (+nmod)`	23	“na manhã desta terça-feira”
	`advmod + case + det + obl`	16	“Desde julho de 2005 até agora”
Attribution	`root (+nsubj) + punct`	42	“O objetivo é prender 17.”
	`case + det + obl 0(+flat:name)`	38	“na Lapa”
	`root + mark`	20	“disse que”
Sequence	`nsubj + root + conj (+obj / obl)`	44	“o diretor assinou um termo de compromisso (...) e que Bruno será transferido para outra unidade”
	`case + det + obl`	32	“às 8h”
	`root + cc + conj`	24	“começou e ficou”
Elaboration	`acl / acl:relcl`	48	“que teria enviado toneladas de cocaína àquele país”
	`case + det + nmod / obl`	34	“na Lapa (zona oeste de São Paulo)”
	`nsubj + cop + root`	18	“a quadrilha de tráfico internacional de drogas e lavagem de dinheiro liderada por Abadia é uma das mais ricas do mundo”
Circumstance	`case + det + obl (+nmod / amod)`	46	“na manhã desta terça-feira”
	`case + det + nmod / obl`	34	“depois do motim”
	`advmod + case + det + obl`	31	“até agora”
	`mark + advcl`	23	“quando Abadia se entregou à polícia colombiana”

Tabela 1 – Distribuição de padrões sintáticos via UD.

A análise das relações UD presentes nas EDUs anotadas como *Sequence* evidencia: (i) coordenação explícita de eventos; (ii) presença de múltiplos núcleos predicativos; (iii) cláusulas adverbiais de progressão; (iv) marcadores discursivos que sinalizam ordenação, frequentemente realizados por advérbios; (v) paralelismo estrutural nominal-verbal; e (vi) complementos circunstanciais. Observa-se incidência quase categórica de coordenação (`cc, conj`), frequentemente associada a múltiplos núcleos predicativos, o que sinaliza a ausência de hierarquia retórica entre as EDUs e sustenta a organização do discurso como sucessão temporal ou procedimental. A recorrência de construções adverbiais e marcadores temporais ancora os eventos em uma progressão cronológica explícita, enquanto o paralelismo estrutural e a repetição de complementos circunstanciais contribuem para a coerência sequencial.

Já a relação *Elaboration* ocorre quando um segmento discursivo satélite acrescenta detalhes, especificações ou exemplificações a uma entidade ou evento descrito no núcleo, sem promover avanço na progressão temática principal. No gênero jornalístico, essa relação é frequentemente realizada por construções sintáticas que introduzem informação acessória de natureza explicativa ou descritiva.

Sob a perspectiva das UD, observam-se: (i) predominância de modificadores nominais e oblíquos, que viabilizam expansões referenciais de localização, quantificação ou caracterização; (ii) uso recorrente de orações relativas e outros modificadores clausais, responsáveis por acrescentar informação a referentes previamente introduzidos; (iii) estruturas adverbiais explicativas ou circunstanciais; (iv) construções apositivas e nomeações, associadas a estratégias jornalísticas de identificação nominal, cargos e nomes próprios; e (v) coordenação empregada como mecanismo de expansão informacional, acumulando atributos ou descrições sem alterar a macroestrutura discursiva.

A relação *Purpose*, na RST, organiza o discurso de modo que um satélite explicita a finalidade, o objetivo ou a intenção de uma ação descrita no núcleo. Em textos jornalísticos, essa relação é comumente realizada por orações adverbiais finais, construções infinitivas, subordinação introduzida por marcadores de finalidade e verbos de ação orientados a complementos que expressam resultado.

A análise dos padrões UD associados à relação *Purpose* revela: (i) marcação explícita de finalidade; (ii) orações adverbiais finais; (iii) modificadores causais de valor final; (iv) complementação infinitiva; (v) objetos sintáticos explícitos da ação final; (vi) complementos circunstanciais associados à finalidade; e (vii) marcadores de intensidade ou delimitação. A elevada quantidade de objetos explícitos e de complementação infinitiva sugere que a finalidade é concebida, nos textos analisados, como uma ação orientada a um resultado claramente identificável. Ademais, a recorrência de modificadores causais (`acl`) indica processos de nominalização de objetivos. Destacam-se ainda combinações frequentes entre relações UD, como `mark+advcl` e `acl+obj`, que apontam para um alinhamento entre sintaxe e discurso nessa relação RST.

A relação *Attribution*, na RST, ocorre quando uma EDU expressa um conteúdo informacional cuja responsabilidade enunciativa é atribuída a uma fonte explícita. Em textos jornalísticos, essa relação é tipicamente caracterizada por verbos de dizer, marcadores de subordinação e menções diretas ou indiretas à fonte da informação.

No plano das UD, observam-se: (i) presença de núcleo predicativo explícito; (ii) marcação formal de atribuição, associada a construções de discurso reportado; (iii) identificação nominal da fonte, reforçando a ancoragem enunciativa; (iv) estruturas nominais que enquadram institucionalmente a fonte, configurando atribuição indireta; e (v) uso de pontuação delimitadora, como vírgulas ou aspas. Mais da metade dos padrões analisados apresenta núcleo predicativo explícito, frequentemente acompanhado de sujeito lexicalizado, o que evidencia a centralidade de atos de dizer, afirmar, declarar ou opinar. Também se observa recorrência de marcadores subordinativos e construções oblíquas introduzidas por preposição, típicas de expressões como “segundo o ministro” ou “de acordo com a assessoria”. Em conjunto, esses resultados apontam para forte convergência entre as relações UD, especialmente `root`, `nsubj`, `mark`, `obl`, `flat:name` e `appos`.

Por fim, a relação *Circumstance*, na RST, expressa informação contextual que enquadra o evento principal em termos de tempo, lugar, modo, condição, duração ou situação, sem promover progressão narrativa. No discurso jornalístico, essa relação é predominantemente realizada por complementos circunstanciais, advérbios e orações adverbiais não finais.

A análise das relações UD evidencia (i) predominância de complementos oblíquos circunstanciais; (ii) modificação adverbial explícita; (iii) orações adverbiais de valor circunstancial; (iv) marcadores subordinantes; e (v) baixa incidência de predicação nuclear. Os resultados indicam uso recorrente de complementos oblíquos e modificadores adverbiais, frequentemente combinados em construções preposicionais múltiplas, características da expressão de informação temporal, espacial ou situacional. A incidência reduzida de núcleos verbais explícitos reforça o caráter acessório dessa relação, que atua majoritariamente como satélite na RST. Além disso, a presença de orações adverbiais e de marcadores subordinativos de não-finalidade confirma que *Circumstance* desempenha a função de fornecer pano de fundo interpretativo ao leitor.

5. Considerações finais

Os resultados obtidos sugerem que os níveis discursivo e sintático, modelados respectivamente pela RST e pela UD, apresentam regularidades sistemáticas em sua interface. Enquanto a UD oferece uma representação estrutural da sintaxe, a RST possibilita observar a organização dessa estrutura em nível macrodiscursivo, evidenciando regularidades na interface entre sintaxe e discurso. Ainda que os achados ainda não permitam estabelecer associações estatisticamente generalizáveis, eles indicam a existência de regularidades promissoras para investigações futuras sobre a interface entre esses dois níveis de análise linguística

A análise comparativa das relações RST investigadas indica que cada uma apresenta uma assinatura sintático-discursiva própria. *Elaboration* associa-se a mecanismos de expansão descritiva; *Attribution* à ancoragem enunciativa explícita; *Sequence* ao encadeamento coordenado de eventos; *Purpose* à marcação formal de finalidade; e *Circumstance* ao fornecimento de enquadramento contextual periférico. Em conjunto, esses resultados corroboram a hipótese de convergência entre padrões de dependência UD e relações discursivas da RST.

Considerando o caráter exploratório deste estudo e o tamanho reduzido da amostra analisada, em nossos trabalhos futuros deveremos ampliar o conjunto de dados, contemplar outras relações RST e diferentes gêneros textuais, bem como incorporar análises estatísticas inferenciais para avaliar a significância das associações observadas entre relações sintáticas e discursivas. Ainda, será importante avaliar a especificidade dos padrões sintáticos identificados em relação às diferentes relações RST, de modo a avaliar a caracterização das relações discursivas, ou ainda quais padrões são compartilhados entre diferentes relações. Ademais, é relevante aprofundar a avaliação das anotações sintáticas geradas automaticamente, especialmente em casos de desalinhamento entre a atribuição da raiz sintática e o estatuto discursivo de satélite em relações RST mononucleares.

Cabe ainda destacar que, como as anotações UD foram produzidas automaticamente pelo PortParser v2, possíveis desvios de análise sintática podem ter influenciado parte dos padrões observados. Apesar de a ferramenta apresentar elevado desempenho para o português, sobretudo para o gênero discursivo analisado, em análises futuras deverá ser feitas avaliações quanto às saídas do *parser* com anotações revisadas manualmente.

Agradecimentos

Este projeto foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei N. 8.248, de 23 de outubro de 1991, no âmbito do PPI-Softex, coordenado pela Softex e publicado como Residência em TIC 13, DOU 01245.010222/2022-44. Além disso, o trabalho contou com suporte financeiro da FAPEB/CNPq da Chamada 004/2023 – Programa Primeiros Projetos.

Referências

Ando, R. K., & Zhang, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research, 6*, 1817–1853. Disponível em: https://www.jmlr.org/papers/volume6/ando05a/ando05a.pdf.

Cardoso, P. C. F., Maziero, E. G., Castro Jorge, M. L. R., Seno, E. M. R., Di Felippo, A., Rino, L. H. M., Nunes, M. G. V., & Pardo, T. A. S. (2011). CSTNews: A discourse-annotated corpus for single- and multi-document summarization of news texts in Brazilian Portuguese. In Proceedings of the 3rd Brazilian RST Meeting (pp. 88–105). Disponível em: https://sites.icmc.usp.br/taspardo/rst2011-cardosoetal1.pdf.

Cardoso, P. C. F., Souza, J. W. da C., & Rodrigues, R. (2024). Modelos discursivos. In H. M. Caseli & M. G. V. Nunes (Eds.), *Processamento de linguagem natural: Conceitos, técnicas e aplicações em português* (3ª ed., Cap. 11). BPLN. Disponível em: https://brasileiraspln.com/livro-pln/3a-edicao/parte-discurso/cap-modelos-discursivos/cap-modelos-discursivos.html.

Cardoso, P., Souza, J., Rodrigues, R., Dantas, E., Santa Bárbara, L., Araújo, M., Gama, N., Almeida, T., & Cruz, G. (2024). A linguagem em foco: Anotação de sinalizadores discursivos em textos jornalísticos. In Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (pp. 247–256). Sociedade Brasileira de Computação. DOI: https://doi.org/10.5753/stil.2024.245329.

Das, D., & Taboada, M. (2014). *RST signalling corpus annotation manual*. Disponível em: https://www.sfu.ca/~mtaboada/docs/publications/RST_Signalling_Corpus_Annotation_Manual.pdf.

Di Felippo, A., *et al.* (2021). Descrição preliminar do Corpus DANTEStocks: Diretrizes de segmentação para anotação segundo Universal Dependencies. In *Proceedings of the XIII Brazilian Symposium in Information and Human Language Technology and Collocated Events* (p. 335). Disponível em: https://aclanthology.org/2021.stil-1.36.pdf.

Di Felippo, A., *et al.* (2024). Genipapo: A multigenre dependency parser for Brazilian Portuguese. In *Proceedings of the Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL 2024)* (pp. 257–266). Sociedade Brasileira de Computação. DOI: https://doi.org/10.5753/stil.2024.245415.

Freitas, I. S. (2025). Descrição da fragmentação estrutural de tweets do mercado financeiro via relação de "parataxis" do modelo Universal Dependencies [Trabalho de conclusão de curso, Universidade Federal de São Carlos]. Disponível em: https://repositorio.ufscar.br/server/api/core/bitstreams/37ec0aca-bfea-473e-a29d-5b579989d69f/content.

Gusfield, D. (1997). *Algorithms on strings, trees and sequences*. Cambridge University Press.

Liu, Z., Shi, K., & Chen, N. (2021). DMRST: A joint framework for document-level multilingual RST discourse segmentation and parsing. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse* (pp. 154–164). DOI: 10.18653/v1/2021.codi-main.15.

Lopes, L., & Pardo, T. (2024). Towards PortParser: A highly accurate parsing system for Brazilian Portuguese following the Universal Dependencies framework. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese (Vol. 1)* (pp. 401–410). Disponível em: https://aclanthology.org/2024.propor-1.41/.

Maekawa, A., Hirao, T., Kamigaito, H., & Okumura, M. (2024). Can we obtain significant success in RST discourse parsing by using Large Language Models? *In* Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2024) (pp. 2803–2815). Association for Computational Linguistics. Disponível em: https://doi.org/10.18653/v1/2024.eacl-long.171.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text, 8*(3), 243–281. Disponível em: https://www.sfu.ca/rst/05bibliographies/bibs/Mann_Thompson_1988.pdf.

Nivre, J. (2015). Towards a universal grammar for natural language processing. In A. Gelbukh (Ed.), *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 3–16). Springer. DOI: https://doi.org/10.1007/978-3-319-18111-0_1.

Rodrigues, R., Souza, J. W. da C., & Cardoso, P. C. F. (2023). Sinalizadores retórico-discursivos: Revisitando a anotação RST no córpus CSTNews. In Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL) (pp. 249–257). Sociedade Brasileira de Computação. DOI: https://doi.org/10.5753/stil.2023.234120.

Rodrigues, R., Souza, J., & Cardoso, P. (2023). Sinalizadores retórico-discursivos: Revisitando a anotação RST no córpus CSTNews. In Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (pp. 249–257). Sociedade Brasileira de Computação. DOI: https://doi.org/10.5753/stil.2023.234120.

Souza, J.W.C., Cardoso, P. & Rodrigues, R. Systematic Review of Studies on Rhetorical Structure Theory (RST) (2023). *Revista De Estudos Da Linguagem*, *31*(3), 1643-1675. DOI: https://doi.org/10.17851/2237-2083.31.3.1643-1675.

Thompson, K., Chaturvedi, A., Hunter, J., & Asher, N. (2024). Llamipa: An Incremental Discourse Parser. *In* Findings of the Association for Computational Linguistics: EMNLP 2024 (pp. 6418–6430). Association for Computational Linguistics. Disponível em: https://doi.org/10.18653/v1/2024.findings-emnlp.373.

Zeldes, A. (2024). *rst2dep* [Computer software]. Disponível em: https://github.com/amir-zeldes/rst2dep. https://pypi.org/project/rst2dep/

Zufferey, S., & Degand, L. (2024). *Connectives and discourse relations*. Cambridge University Press.

Classificação de Textos com Matrizes de Similaridade: Um Estudo de Caso em um Corpus com Nomes de Produtos

Antonio Ferreira de Castro Barbosa[1], Eduardo Corrêa Gonçalves[2]

[1]Bacharelado em Sistemas de Informação – Universidade Federal do Estado do Rio de Janeiro (BSI/UNIRIO)
Avenida Pasteur, 458 – 22.290-255 – Rio de Janeiro – RJ – Brasil

[2]Escola Nacional de Ciências Estatísticas (ENCE/IBGE)
Rua André Cavalcanti, 106 – 20.231-050 – Rio de Janeiro – RJ – Brasil

antoniofdcbarbosa@gmail.com, eduardo.correa@ibge.gov.br

Abstract. *This paper addresses the problem of semantic comparison of pairs of short strings in Portuguese. More specifically, this work evaluates an unsupervised approach based on similarity matrices for classifying product names that may contain different specifications. The matrices were evaluated individually and combined in pairs. In preliminary experiments on a database of 1,000 instances, combining the TF-IDF matrix with matrices built with pre-trained BERT embeddings achieved the best accuracy.*

Resumo. *Este trabalho aborda o problema da comparação semântica de pares de strings curtas em português. Mais especificamente, o trabalho avalia uma abordagem não supervisionada baseada em matrizes de similaridade para classificar nomes de produtos contendo diferentes tipos de especificações. As matrizes foram utilizadas de forma isolada e combinadas aos pares. Em experimentos realizados sobre um corpus composto por 1.000 instâncias, a melhor acurácia foi obtida através da combinação da matriz TF-IDF com matrizes de embeddings pré-treinados da família BERT.*

1. Introdução

Plataformas digitais usualmente empregam técnicas de busca semântica para aprimorar a experiência de busca em suas aplicações [Huang et al. 2020]. Por exemplo, em uma aplicação de busca de produtos em um supermercado, é possível contornar o problema da falta de resultados com duas possibilidades diferentes: (i) apresentar o mesmo produto, mas de marcas alternativas (ex.: o cliente procura por manteiga da marca A, que não existe no estoque, então a aplicação retorna as manteigas das marcas B e C); (ii) apresentar produtos similares, independente da marca (nesse caso, a aplicação poderia retornar uma margarina). Para que o sistema seja capaz de apresentar esses resultados similares, é necessário que haja uma forma de classificar automaticamente os nomes comerciais dos produtos para nomes padronizados [Köpcke and Rahm, 2010].

Nesse contexto, o presente trabalho utiliza-se de um corpus de 1.000 pares relacionando nomes de produtos alimentícios vendidos em supermercados com o seu nome padronizado no Índice Nacional de Preços ao Consumidor (INPC) do IBGE [IBGE 2016]. A partir desse *dataset*, matrizes de similaridade de cosseno para diferentes representações dos textos (TF-IDF [Manning et al. 20028], BERTimbau Base

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 540–545
October 19-22, 2026 ©2026 Association for Computational Linguistics

[Souza et al. 2020], mBERT [Google 2020] e Universal Sentence Encoder (USE) [Cer et al. 2018]) foram utilizadas para classificar o nome INPC correto de cada produto.

2. Trabalhos Relacionados e Metodologia

2.1. Trabalhos Relacionados

Romualdo et al. (2021) abordam a tarefa de mensurar o grau de similaridade entre títulos de produtos comercializados em um grande site de *e-commerce*. Tais títulos são formados não apenas pelo nome do produto como também por uma série de especificações (ex.: "Cartucho Epson 196 Preto 5ml T196120"). Em testes com um dataset de 400 pares de títulos, *embeddings* genéricos mBERT e BERTimbau superaram os estáticos treinados com textos do domínio (FastText, Word2Vec e GloVe [Jurafsky e Martin 2026]) ao distinguir produtos similares de dissimilares.

Araujo et al. (2023) apresentam um estudo de caso que investigou o emprego de matrizes de similaridade de cosseno [Manning et al. 2008] para identificar boletins policiais semelhantes. Diferentes representações foram avaliadas, tais como TF-IDF e *embeddings* de sentenças USE e SBert [Reimers and Gurevych 2019] pré-treinados e treinados com textos do domínio. Em uma coleção de 1.089 boletins anotados, o melhor resultado foi obtido pela representação TF-IDF (77,38% de acurácia), seguida pelos embeddings de sentença pré-treinados USE (70,47%).

Meirelles et al. (2022) propõem o uso combinado de matrizes de similaridade para considerar simultaneamente medidas que atuam em diferentes subáreas do sistema linguístico [Caseli et al. 2026]. A avaliação utilizou uma base com 3.910 pares casados de nomes de produtos e serviços de duas pesquisas do IBGE (POF e INPC) [IBGE 2026]. O estudo constatou que combinar medidas que atuam no nível morfológico, como Jaro-Winkler [Winkler at al.1990], com uma medida semântica (cosseno de *embeddings* Word2Vec pré-treinados do NILC [Hartmann et al. 2017]) aumentou a eficácia do processo de pareamento de nomes equivalentes.

2.2. Base de Dados

O presente estudo utiliza um corpus com 1.000 pares (x, y) de nomes de produtos alimentícios [de Lima and Gonçalves, 2022], em que x representa o nome de um produto vendido em um supermercado, enquanto y representa a classe (nome padronizado) do produto no INPC [IBGE, 2016]. Há um total de 470 possíveis classes. Alguns exemplos de pares (x, y) do corpus são: {(*Repolho Verde (unidade) – orgânico*, *Repolho orgânico*), (*Patê de Atum Gomes da Costa*, *Patê*), (*Carre Suino Congelado Pc 3,5 Kg*, *Carne de porco*)}. Os nomes dos produtos nos supermercados são mais longos (34,4 caracteres e 6,5 tokens em média) e podem conter erros ortográficos (ex.: "Carre" em vez de "Carré"), além de abreviações, nomes de marcas, unidades de medida e outras especificações. Por sua vez, os nomes do INPC são padronizados e concisos (média de 13,5 caracteres e 2,2 tokens).

2.3. Matrizes de Similaridade

A partir do corpus descrito acima, torna-se possível empregar uma abordagem supervisionada para treinar um modelo que classifica o nome de um produto de supermercado em um nome do INPC. Contudo, uma dificuldade para esta abordagem é que o *dataset* contém, em média, apenas 2,13 instâncias por classe (são 1.000 instâncias, cada qual anotada com uma dentre 470 classes possíveis). Por esta razão, decidiu-se

tratar o problema utilizando uma estratégia não supervisionada baseada em matrizes de similaridade. Cada matriz de similaridade *M* avaliada possui 1.000 linhas (nomes dos produtos nos supermercados) e 470 colunas (nomes do INPC, que representam as classes do problema). As células representam o valor da similaridade de cosseno entre cada nome das linhas e cada nome das colunas de acordo com alguma representação vetorial. Para os experimentos preliminares apresentados na próxima seção, foram escolhidas as representações com melhor desempenho nos trabalhos de Araujo et al. (2023) e Romualdo et al. (2021): TF-IDF, embeddings pré-treinados de palavras BERTimbau Base e mBERT e embeddings pré-treinados de sentença USE. Foram avaliadas matrizes de similaridade individuais dessas representações e combinadas aos pares, de acordo com a proposta de Meirelles et al. (2022) (detalhada na Seção 3.2).

3. Experimentos

Nesta seção, são apresentados os resultados obtidos nos experimentos. O pacote *scikit-learn* [Pedregosa et al. 2011] foi utilizado para obter a representação TF-IDF dos textos, enquanto as implementações disponibilizadas no *Hugging Face* [Hugging Face 2026] foram utilizadas para os modelos BERTimbau, mBERT e USE, que não foram submetidos a ajuste fino. Os modelos foram testados com e sem o pré-processamento das strings, que consistiu na remoção de stop words e lematização com a ferramenta *spaCy* [Honnibal and Montani 2017]. Para TF-IDF, BERTimbau e mBERT, os melhores resultados ocorreram com pré-processamento; já para o USE, o desempenho foi superior sem essa operação. Dois tipos de métricas de avaliação de desempenho foram adotadas. A primeira é a Acurácia Estrita [Meirelles et al. 2022], que consiste na proporção de vezes em que a matriz de similaridade mapeou única e corretamente a classe INPC para um nome de produto. A segunda é a Acurácia *Top-k*, que considera a classificação correta caso o valor de similaridade da classe real do produto esteja entre os *k* maiores da matriz.

3.1. Desempenho Isolado das Matrizes

A Tabela 1 apresenta os resultados do primeiro experimento, que comparou o desempenho individual das matrizes. A representação TF-IDF obteve Acurácia Estrita de 61,90%, quase 25% superior ao do segundo melhor resultado, obtido pelo BERTimbau. O TF-IDF também alcançou maiores valores para a Acurácia *Top-k*, mas dessa vez com menor diferença para os embeddings (ex.: diferença de 7,2% para o BERTimbau na Acurácia *Top*-10). É possível ainda observar que o crescimento no valor da Acurácia Estrita para a Acurácia *Top-k* é menos acentuado no TF-IDF do que nos embeddings. Isso pode ser justificado pelo fato de a matriz TF-IDF ser muito esparsa (média de apenas 7,1 células com valores diferentes de zero por linha).

Tabela 1. Desempenho das Matrizes Individuais

Matriz de Similaridade	Acurácia Estrita	Acurácia *Top*-5	Acurácia *Top*-10
TF-IDF	**0,6190**	**0,6820**	**0,6890**
BERTimbau	0,3700	0,5640	0,6170
mBERT	0,3530	0,5190	0,5870
USE	0,2650	0,4380	0,5250

O resultado dos *embeddings* pode estar relacionado ao tamanho muito curto das strings e a consequente falta de contexto. Por exemplo, a representação USE (pior valor de Acurácia Estrita, com 26,50%) foi projetada para codificar sentenças. Porém, no experimento foi empregada para representar vetorialmente nomes muito curtos, frequentemente formados apenas por um substantivo principal seguido de um reduzido número de adjetivos ou locuções adjetivas (ex.: "Rabada bovina congelada 1Kg").

3.2. Desempenho de Pares de Matrizes Combinadas

O segundo experimento avaliou pares de matrizes combinadas, utilizando a abordagem proposta em Meirelles et al. (2022). Nesta abordagem, uma matriz híbrida M_H é obtida utilizando a Equação 1, em que M_1, M_2, ..., M_n são *n* matrizes de similaridade previamente geradas, enquanto $\alpha \in \mathbb{R}+$ atua como ponderação, possibilitando dar maior peso para valores de similaridade maiores. Para o experimento, foi utilizado $\alpha = 1$.

$$M_H = \frac{(M_1^\alpha + M_2^\alpha + \cdots + M_n^\alpha)}{n} \quad (1)$$

Foram geradas 3 matrizes híbridas M_{H1}, M_{H2} e M_{H3} representando, respectivamente, a média da matriz TF-IDF (melhor resultado no experimento anterior) com as matrizes BERTimbau, mBERT e USE. Os resultados são apresentados na Tabela 2. Com relação à Acurácia Estrita, as matrizes híbridas MH_1 e MH_2 (TF-IDF e BERTimbau e TF-IDF e mBERT) obtiveram resultados ligeiramente superiores a melhor matriz do experimento anterior (TF-IDF). Contudo, o mais interessante foi a grande melhora nos valores da Acurácias *Top-k* das matrizes combinadas. Considerando a aplicação descrita na Introdução, um sistema de busca de produtos em supermercados, o resultado de Acurácia *Top*-5 acima de 81% obtido pelas matrizes híbridas indica que, na maioria das vezes, o nome INPC correto estaria entre as primeiras sugestões apresentadas ao usuário. Esse resultado é especialmente relevante em cenários em que o sistema não necessita de uma classificação única e objetiva, mas sim de uma seleção de alternativas razoáveis, como na substituição de produtos fora de estoque.

Tabela 2. Desempenho das Matrizes Combinadas

Matriz de Similaridade	Acurácia Estrita	Acurácia Top-5	Acurácia Top-10
MH_1 (TF-IDF e BERTimbau)	**0,6290**	0,8180	0,8480
MH_2 (TF-IDF e mBERT)	0,6280	0,8190	0,8450
MH_3 (TF-IDF e USE)	0,5980	**0,8230**	**0,8610**

4. Conclusões

Os resultados dos experimentos preliminares sugerem que a combinação de matrizes de similaridade pode aprimorar o processo de classificação e busca semântica de textos curtos, como nomes de produtos. Entretanto, no corpus testado, os resultados dos *embeddings* pré-treinados isoladamente foram bem inferiores aos do TF-IDF. Sendo assim, no futuro pretende-se realizar o ajuste fino desses modelos utilizando bases de dados públicas que contenham nomes de produtos. Intenciona-se ainda avaliar o desempenho de modelos *decoder-only* (como em Pellegrini et al. (2025) e Yepez et al. (2024)) e empregar outras estratégias para combinar as matrizes de similaridade.

5. Limitações do Estudo

Os resultados apresentados nesse estudo são preliminares e envolvem um único conjunto de dados composto por 1.000 instâncias. Foram utilizadas versões pré-treinadas dos embeddings avaliados, sem a realização de ajuste fino com textos do domínio.

6. Declaração de Ética

O corpus utilizado foi construído exclusivamente a partir de dados de acesso público, sendo eles os nomes dos produtos do INPC do IBGE e nomes de produtos anunciados em sites de supermercados. Desta forma, não foi necessária a aprovação por parte de comitê de ética. O método proposto pode ser visto como uma ferramenta de apoio que visa aprimorar o desempenho de sistemas de busca semântica.

7. Uso de IA Generativa

Os autores declaram que houve utilização da IA generativa somente como apoio à implementação de trechos do código utilizado nos experimentos. Não houve a utilização de IA generativa para produzir ou modificar o texto do artigo.

Referências

Araújo, J., Silva, T., Rocha, A., and Lira, V. (2023). Police Report Similarity Search: A Case Study. In *Proceedings of the XII Brazilian Conference on Intelligent Systems (BRACIS 2023)*, pages 394-409, Belo Horizonte, Springer.

Caseli, H. M., Nunes, M. G. V., and Pagano, A. (2026). O que é pln? In Caseli, H. M. and Nunes, M. G. V., editors, Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português – Volume 1, book chapter 1. BPLN, 4 ed.

Cer, D., et al. (2018). Universal sequence encoder. arXiv preprint arXiv:1803.11175. Disponível em: https://arxiv.org/abs/1803.11175.

de Lima, L. S. G. and Gonçalves, E. C. (2022). Similaridade Semântica de Nomes de Produtos Alimentícios Utilizando Wordnets do Português. In *Proceedings of the 15th Seminar on Ontology Research in Brazil (ONTOBRAS) and 6th Doctoral and Masters Consortium on Ontologies (WTDO)*, Online, CEUR-WS.org.

Google (2020). BERT. Disponível em: https://github.com/google-research/bert

Hartmann, N. S., et al. (2017). Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology (STIL 2017)*, pages 122–131, Uberlândia, SBC.

Honnibal, M. and Montani, I. (2017). "spaCy: Industrial-strength Natural Language Processing in Python", https://spacy.io.

Huang, J.-T., Sharma, A., Sun, S., Xia, X., Zhang, D., et al. (2020). Embedding-based Retrieval in Facebook Search. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2553–256, Online. ACM.

Hugging Face. (2026). "Models", https://huggingface.co/

IBGE (2016). Para compreender o INPC (um texto simplificado), 7a. ed, IBGE.

IBGE (2026). “Inflação”, https://www.ibge.gov.br/explica/inflacao.php.

Jurafsky, D. and Martin, J. H. (2026) “Embeddings”. In: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models. 3rd ed. Stanford University, 2026. Cap. 5. Disponível em: https://web.stanford.edu/~jurafsky/slp3/. Acesso em: 14 mai. 2026.

Köpcke, H. and Rahm, E. (2010). Frameworks for entity matching: A comparison. *Data & Knowledge Engineering*, 69(2): 197–210. Elsevier

Manning, C. D., Raghavan, P., and Schütze, H. (2008). Introduction to information retrieval, Cambridge University Press.

Meirelles, T. P., Gonçalves, E. C., and Gomes, D. T. (2022). Pareamento de Nomes de Produtos e Serviços Utilizando Medidas de Similaridade Textual nos Níveis Alfabético, Léxico e Semântico. *Cadernos Do IME - Série Informática*, 46:104–117. IME / UERJ.

Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830. MIT Press.

Pellegrini, L., Santos, F., and Cantarino, F. (2025). Classificação de Notícias Falsas na Língua Portuguesa Utilizando Modelos baseados na Arquitetura Transformer. In *Proceedings of the 16th Brazilian Symposium in Information and Human Language Technology (STIL 2025)*, pages 549-556, Fortaleza. SBC.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084. Disponível em: https://arxiv.org/abs/1908.10084.

Romualdo, A., Real, L., and Caseli, H. (2021). Measuring Brazilian Portuguese Product Titles Similarity Using Embeddings. In *Proceedings of the 13th Brazilian Symposium in Information and Human Language Technology (STIL 2021)*, pages 121-132, Online. ACL.

Souza, F., Nogueira, R. and Lotufo, R. (2020). BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Proceedings of the 2020 Brazilian Conference on Intelligent Systems (BRACIS 2020)*, pages 403–417, Online. Springer.

Winkler, W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage”. In *Proceedings of the Sect. on Surv. Research*, pages 354–359, ERIC.

Yepez, J., Tavares, B., Peres, F., and Becker, K. (2024). Na Batida do Funk: Modelagem de Tópicos Combinando LLM, Engenharia de Prompt e BERTopic. *In Anais do XXXIX Simpósio Brasileiro de Bancos de Dados* (SBBD 2024), pages 613–625, Florianópolis, SBC.

Uso de modelos de linguagem na identificação de relações de coerência da *Rhetorical Structure Theory* em textos de *User-Generated Content*

Pedro Moreno Niella de Souza Sales Evangelista[1], Jackson Wilke da Cruz Souza[2]

[1]Instituto de Ciência, Tecnologia e Inovação - Universidade Federal da Bahia (UFBA)
Camaçari, Bahia / Brasil

[2]Programa de Pós-Graduação em Língua e Cultura - Universidade Federal da Bahia
Salvador, Bahia / Brasil

{pedroniella, jackcruzsouza}@gmail.com

Abstract. *This work investigates the automatic identification of Rhetorical Structure Theory (RST) rhetorical relations in User-Generated Content, using Discourse Signals (DSs) as coherence evidence under the RST framework. The low coverage of classical connectives (19%) motivates a comparative pipeline with lexical heuristics, Support Vector Machine, and BERTimbau fine-tuning on the DANTEStocks corpus (4,048 financial tweets). BERTimbau achieved 95% accuracy and 93% weighted F1-score; however, class imbalance restricts the macro F1-score to 53%, showing that deep contextual modeling mitigates UGC linguistic noise but does not resolve the class imbalance problem.*

Resumo. *Este trabalho investiga a identificação automática de relações retóricas Rhetorical Structure Theory (RST) em User-Generated Content, utilizando Sinalizadores Discursivos (SDs) como evidências de coerência sob a ótica da RST. A baixa cobertura de conectivos clássicos (19%) motiva um pipeline comparativo com heurísticas léxicas, Support Vector Machine e fine-tuning do BERTimbau no corpus DANTEStocks (4.048 tweets financeiros). O BERTimbau alcançou 95% de acurácia e 93% de F1-Score ponderado; contudo, o desbalanceamento entre classes restringe o F1-Macro a 53%, evidenciando que a modelagem contextual profunda mitiga o ruído linguístico do UGC, mas não resolve o problema de desbalanceamento.*

1. Introdução

User-Generated Content (UGC) constitui uma das principais fontes de dados textuais de interesse para o Processamento de Linguagem Natural (PLN) [Gama e Souza, 2024]. Contudo, sua análise computacional esbarra em obstáculos como desvios linguísticos, variações ortográficas e fragmentação sintática, típicos de publicações em redes sociais [Di-Felippo et al., 2021]. No domínio financeiro, esses desafios se agravam pela brevidade das postagens e por tokens específicos, exigindo metodologias que superem a identificação de palavras-chave e alcancem a organização pragmático-discursiva das sentenças.

A *Rhetorical Structure Theory* (RST) [Mann e Thompson, 1988] oferece o suporte teórico necessário: a teoria decompõe o texto em *Elementary Discourse Units* (EDUs) e organiza as relações de coerência entre elas em estrutura hierárquica. Até o momento, a identificação dessas relações é majoritariamente feita com base em Marcadores Discursivos (MDs), especialmente, preposições e conjunções. Contudo, Das e Taboada (2019) demonstraram que a coerência textual (i) pode ocorrer a despeito de MDs comuns,

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 546–550
October 19-22, 2026 ©2026 Association for Computational Linguistics

e (ii) sempre deixa evidências das relações no enunciado. Como resultado, os autores propuseram uma matriz de Sinalizadores Discursivos (SDs) que ampliam a noção de pistas de identificação das relações de coerência, considerando outros aspectos importantes para a sinalização dessas relações, como pontuação.

No cenário da língua portuguesa, esforços recentes têm mapeado a ocorrência desses sinalizadores em diferentes gêneros. Cardoso *et al.* (2024) e Gama e Souza (2024) sistematizaram a anotação RST para os domínios jornalístico e financeiro (este advindo de postagens de redes sociais), respectivamente, resultando em uma tipologia de SDs que indicam relações de coerência.

A automatização de algumas tarefas em PLN evoluiu de sistemas heurísticos rígidos para abordagens estatísticas e redes neurais. Conforme discutido por Monard e Baranauskas (2003), classificadores supervisionados tradicionais oferecem robustez ao operar sobre espaços vetoriais definidos por atributos extraídos de esquemas formais de anotação linguística [Leech, 1993]. O estado da arte foi impulsionado pelo surgimento de modelos baseados em *Transformers*, como o BERTimbau [Souza, Nogueira e Lotufo, 2020], capaz de capturar nuances sintáticas e semânticas contextuais.

Neste trabalho, propusemos e avaliamos um *pipeline* comparativo estruturado em três abordagens distintas: heurística, Aprendizagem de Máquina (AM) supervisionado clássico e a aplicação do modelo BERTimbau. O objetivo deste estudo é investigar em que medida a incorporação de SDs e de representações contextuais profundas permite classificar automaticamente relações RST em *tweets* financeiros.

Este estudo busca somar-se aos esforços históricos de análise discursiva automática para o português, como o *corpus* CSTNews [referência] e do *parser* discursivo DiZer [Pardo e Seno, 2005]. Como contribuição inovativa, apresentamos a expansão do escopo dos trabalhos para o domínio desafiador do UGC financeiro por meio de arquiteturas baseadas em *Transformers*.

2. Metodologia

O *pipeline* computacional para identificação e classificação de relações retóricas foi estruturado em ambiente Python, com as bibliotecas Scikit-Learn [Pedregosa *et al.,* 2011] e Hugging Face Transformers [Wolf *et al.,* 2020], em três abordagens: heurísticas léxicas, classificação supervisionada clássica e ajuste fino do modelo BERTimbau, aplicado ao *corpus* DANTEStocks. A variável-alvo compreende 12 relações RST mapeadas via rotulação semiautomática: cada sentença foi revisada em relação à tipologia de SDs proveniente do trabalho de Cardoso *et al.* (2024), e a relação foi atribuída conforme o conectivo detectado; sentenças sem SDs receberam o rótulo default *Elaboration*, seguindo a própria proposta de Mann e Thompson (1988). Essas 12 relações foram as únicas com ocorrências identificáveis via SDs explícitos no *corpus*. Embora os MDs tenham ocorrido em apenas cerca de 19% do *corpus*, a inclusão de sinalizadores gráficos e semânticos (como sinais de pontuação e menções de usuários) no *pipeline* heurístico elevou significativamente a cobertura, o que justifica a relação *Preparation* figurar como a classe majoritária, como demonstra-se na Tabela 1. Sentenças sem nenhum sinalizador detectado receberam o rótulo *default Elaboration*.

Na segunda abordagem, utilizamos o algoritmo *Support Vector Machine* (SVM) [Cortes e Vapnik, 1995] com *kernel* RBF para modelar uma matriz atributo-valor de baixa dimensão a partir dos *tweets*. Os 23 atributos foram organizados em quatro grupos conforme sua natureza linguística e funcional. Essa distinção é útil para mensurar, quais dimensões mais contribuem para a predição das relações RST. Os grupos são divididos

em *morfossintáticos absolutos* que são a contagem de tokens para 11 etiquetas advindas da *Universal Dependencies* [Nivre *et al.,* 2016]; *proporcionais, sendo valores relativos, calculados pela* razão entre a frequência de uma categorias críticas pelo tamanho da sentença; *morfológicos binários são* valores lógicos que indicam exclusivamente a presença ou ausência de flexões verbais (pretérito ou futuro) no campo *Feats*; e *estruturais/específicos cujo contêm* características próprias do domínio financeiro e da rede social, incluindo a posição relativa da sentença e a presença de hashtags, menções, URLs, tickers, percentuais e vocabulário da área. Para o AM, utilizamos validação cruzada de 5 pastas, com o hiperparâmetro `class_weight='balanced'` para lidar com o desbalanceamento de classes. Para o custo e a largura de banda do *kernel*, assumiu-se `C = 1.0` e `gamma = 'scale'`. Todos os outros hiperparâmetros foram mantidos *default* da biblioteca *scikit-learn*. A distribuição das relações RST neste estudo está da seguinte maneira: *Preparation* com 2.210 instâncias, *Elaboration* com 651, *Circumstance* com 451, *Condition* com 404, *Cause* com 85, *Contrast* com 84, *Concession* com 48, *List* com 41 e *Antithesis*, *Interpretation, Restatement, Background e Purpose com 74*

Como limite de desempenho, realizou-se o ajuste fino do modelo BERTimbau Base com 110 milhões de parâmetros. O tokenizador nativo processou os tweets com truncamento e *padding* limitados a 128 *tokens*. O dataset, então, foi particionado em 80% para treino (3.238 instâncias) e 20% para teste (810 instâncias). O treinamento ocorreu com GPU ao longo de 5 épocas, utilizando o otimizador AdamW com taxa de aprendizagem de $2x10^{-5}$. Além disso, foram definidos os seguintes hiperparâmetros para a calibração do modelo: `batch size` de 32, `weight decay` de 0.01, `warmup ratio` de 0.1 e `seed` igual a 42; para a avaliação final, baseamo-nos no conjunto de teste isolado.

3. Resultados

A avaliação contrastiva entre as abordagens evidencia o impacto da modelagem contextual em ambientes de UGC. Enquanto a abordagem heurística restringiu-se à presença explícita de SDs (cobertura de 19%), os modelos supervisionados estendem a classificação para a totalidade do *corpus*. O SVM alcançou acurácia total de 64% e F1-Macro de 31% na validação cruzada. O ajuste fino do BERTimbau estabeleceu o limite de desempenho do *pipeline*, atingindo 95% de acurácia total e 93% de *F1-Score* ponderado no conjunto de testes.

Com relação à análise de desempenho das classes de alta frequência, isto é, relações RST com maior número de instâncias no *corpus*, o BERTimbau concentrou-se nessas classes de maior suporte. A abundância de instâncias viabilizou a convergência do modelo, mas evidencia a necessidade de ampliar dados para relações subrepresentadas. A relação *Preparation* (442 instâncias) apresentou comportamento quase invariável, com Precisão de 98%, Revocação de 99% e *F1-Score* de 99%. Algo semelhante ocorreu com as relações *Circumstance* (*F1-Score* de 97% a partir de 90 instâncias), *Condition* (F1: 94%; 81 instâncias) e *Elaboration* (F1: 94%; 130 instâncias). Esse resultado reflete a regularidade dos SDs sintáticos associados a essas relações: conjunções como se e quando funcionam como âncoras de aprendizado, fornecendo ao BERTimbau pistas contextuais suficientes para distingui-las com alta confiança.

Quanto ao diagnóstico de erros nas classes de cauda longa, fenômeno em que poucas classes concentram a maior parte das instâncias enquanto muitas classes possuem frequência muito baixa [Buda, Maki e Mazurowski, 2018], o modelo apresentou

decréscimo ao lidar com relações retóricas de baixa frequência, resultando em um F1-Macro total a 53%. No teste, as relações *Antithesis* e *Interpretation* (ambas com apenas 4 instâncias cada) zeraram em todas as métricas de avaliação automática. Já as relações *Concession* e *List*, com 10 e 8 instâncias respectivamente, apresentaram *F1-Score* de 18% e 20%, sugerindo baixa eficácia para essas classes. Esse resultado é explicado, sobretudo, pela baixa quantidade de exemplos no *corpus*, já que relações RST com menos de dez ocorrências não fornecem variabilidade linguística suficiente para distingui-las com exatidão. Consequentemente, o classificador colapsou as predições raras em direção às classes majoritárias, a saber *Elaboration* e *Preparation*.

4. Considerações finais

Neste trabalho avaliamos a identificação automática de SDs no *corpus* DANTEStocks sob o escopo teórico da RST, contrastando heurísticas lexicais, SVM e o modelo BERTimbau. Os resultados dos experimentos sugerem que dicionários de conectivos são insuficientes para UGC. Além disso, o ajuste fino do BERTimbau demonstrou que arquiteturas de atenção conseguem generalizar além dos MDs explícitos, capturando padrões discursivos que escapam a abordagens léxicas. Ainda, o desbalanceamento quantitativo entre as relações RST restringiu o F1-Macro global, colapsando as categorias minoritárias.

Cabe destacar a circularidade metodológica inerente a este estudo preliminar. O fato de o BERTimbau ter sido treinado em um *corpus* com rotulação semiautomática (*silver-standard*) estritamente baseada nos SDs, a alta acurácia (95%), atesta-se a capacidade do modelo em aprender e replicar as regras heurísticas, e não necessariamente a concordância com a anotação que o humano faria (tida como *gold-standard*). Assim, a validação manual de uma amostra para aferir a real generalização do modelo será uma etapa essencial prevista para trabalhos futuros.

Ainda sobre trabalhos futuros, destacamos a necessidade de investigar métodos para a mitigação desse cenário de cauda longa, englobando técnicas de aumento de dados textuais (*data augmentation*), funções de perda sensíveis ao custo (*cost-sensitive learning*) e estratégias de *few-shot learning* via modelos de linguagem de grande escala. Ainda, pretendemos, de igual modo, aplicar o *pipeline* em outros *corpora* de UGC para validar a capacidade de generalização dos classificadores discursivos em múltiplos domínios informais da *Web.* Por fim, caberá avaliações qualitativas para mensurar linguisticamente a amplitude e profundidade dos resultados obtidos neste estudo preliminar.

Agradecimentos

Este projeto foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei N. 8.248, de 23 de outubro de 1991, no âmbito do PPI-Softex, coordenado pela Softex e publicado como Residência em TIC 13, DOU 01245.010222/2022-44. Além disso, tivemos apoio de recursos financeiros da Chamada FAPESB/CNPq 004/2023 – Programa Primeiros Projetos.

Referências

Buda, M., Maki, A. and Mazurowski, M. A. (2018) "A systematic study of the class imbalance problem in convolutional neural networks", *Neural Networks*, v. 106, p. 249-259.

Cardoso, P., Souza, J., Rodrigues, R., Dantas, E., Bárbara, L. S., Araújo, M., Gama, N., Almeida, T., Cruz, G. (2024) "A Linguagem em Foco: Anotação de Sinalizadores Discursivos em Textos Jornalísticos", STIL 2024, p. 247–256.

Cortes, C. and Vapnik, V. (1995) "Support-vector networks", *Machine Learning*, v. 20, n. 3, p. 273-297.

Das, D. and Taboada, M. (2019) "Multiple signals of coherence relations", *Discours*, v. 24, p. 1-38.

Di-Felippo, A. *et al.* (2021) "Descrição preliminar do corpus DANTEStocks", In: *Anais do XIII Brazilian Symposium in Information and Human Language Technology (STIL 2021)*, p. 210-214.

Gama, N. S. e Souza, J. W. C. (2024) "Estudo preliminar sobre sinalizadores discursivos para Conteúdo Gerado por Usuário", STIL 2024, p. 418–423.

Leech, G. (1993) "Corpus annotation schemes", *Literary and Linguistic Computing*, v. 8, n. 4, p. 275-281.

Mann, W. C. and Thompson, S. A. (1988) "Rhetorical Structure Theory: Toward a functional theory of text organization", *Text & Talk*, v. 8, n. 3, p. 243-281.

Monard, M. C. and Baranauskas, J. A. (2003) "Conceitos sobre aprendizado de máquina", In: *Sistemas Inteligentes: Aplicações a Modelos Biológicos e Biomédicos*, p. 89-114.

Nivre, J. *et al.* (2016) "Universal Dependencies v1: A Multilingual Treebank Collection", In: *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, p. 1659-1666.

Pedregosa, F. *et al.* (2011) "Scikit-learn: Machine Learning in Python", Journal of Machine Learning Research, v. 12, p. 2825-2830.

Souza, F., Nogueira, R. and Lotufo, R. (2020) "BERTimbau: Pretrained BERT Models for Brazilian Portuguese", In: *Proceedings of the 9th Brazilian Conference on Intelligent Systems (BRACIS 2020)*, p. 405-417.

Wolf, T. *et al.* (2020) "Transformers: State-of-the-Art Natural Language Processing", In: Proceedings of EMNLP 2020: System Demonstrations, p. 38-45.

Expansão da Ferramenta Doccano para a Anotação de Análise de Sentimento Estruturada

Antonio Paulo Martins Ferri[1], **Guilherme Dallmann Lima**[1], **Larissa Astrogildo de Freitas**[1], **Andrew Campos**[1]

[1]Universidade Federal de Pelotas (UFPel)
Pelotas – RS – Brasil

apmferri@inf.ufpel.edu.br, gdlima@inf.ufpel.edu.br
larissa@inf.ufpel.edu.br, abdcampos@inf.ufpel.edu.br

Abstract. *Annotation tools are essential for building datasets in Natural Language Processing (NLP). However, the Doccano version adapted for Aspect-Based Sentiment Analysis (ABSA) does not explicitly represent the author of an opinion. This work presents an extension that incorporates the* opinion holder *into the annotation structure, modifying the data model, import/export workflows, and annotation interface. A preliminary evaluation annotated 30 hotel reviews in both versions: 10 contained an explicitly identifiable opinion holder, cases in which the extension enabled a representation unavailable in the baseline version.*

Resumo. *Ferramentas de anotação são fundamentais para a construção de conjuntos de dados em Processamento de Linguagem Natural (PLN). Contudo, a versão do Doccano adaptada para Análise de Sentimento Baseada em Aspectos (ASBA) não representa explicitamente o autor de uma opinião. Este trabalho apresenta uma extensão dessa ferramenta que incorpora o opinion holder à estrutura de anotação, modificando o modelo de dados, os fluxos de importação e exportação e a interface. A avaliação preliminar anotou 30 avaliações de hotéis em ambas as versões: 10 continham um autor explicitamente identificável, casos em que a extensão permitiu registrar uma associação indisponível na versão base.*

1. Introdução

A construção de conjuntos de dados anotados constitui uma etapa importante no desenvolvimento e avaliação de modelos de Processamento de Linguagem Natural (PLN). Nesse contexto, ferramentas como o Doccano [Nakayama et al. 2018] auxiliam na anotação de textos para diferentes tarefas, incluindo análise de sentimentos.

Nesse cenário, a Análise de Sentimento Baseada em Aspectos (ASBA) busca identificar aspectos mencionados em um texto e os sentimentos associados a eles. Tarefas de ASBA frequentemente envolvem subtarefas como Extração de Aspecto e Classificação de Sentimento do Aspecto [Pontiki et al. 2016]. Embora a literatura de mineração de opiniões reconheça o autor da opinião (*opinion holder*) como um elemento relevante da estrutura opinativa [Liu 2012], sua representação explícita não está necessariamente presente em esquemas e ferramentas voltados à ASBA. A Análise de Sentimento Estruturada (ASE), por sua vez, inclui o emissor da opinião e sua expressão de sentimento [Barnes et al. 2021], ampliando a informação a respeito da opinião.

O objetivo deste trabalho é expandir a versão do Doccano previamente adaptada para ASBA [Cardozo 2025], incorporando a representação explícita do autor da opinião. Para isso, foi introduzido o *span* "autor" e sua associação às relações entre aspecto e opinião. As principais contribuições do trabalho são: (i) extensão do modelo de dados

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 551–555
October 19-22, 2026 ©2026 Association for Computational Linguistics

para representar autoria; (ii) adaptação dos fluxos de importação e exportação; e (iii) modificação da interface para criação e visualização dessas relações.

A representação explícita do autor permite distinguir situações em que diferentes indivíduos expressam opiniões sobre um mesmo aspecto [Wiebe et al. 2005]. Dessa forma, a estrutura anotada passa a preservar não apenas o conteúdo opinativo e sua polaridade, mas também a associação entre a opinião e seu respectivo emissor.

2. Metodologia

2.1. Limitações da Extensão do Doccano para ASBA

A extensão do Doccano para ASBA permite a marcação de aspectos e opiniões, representados, respectivamente, pelos *spans* "aspecto" e "opinião", relacionados pelas polaridades "positiva", "negativa" e "neutra". Entretanto, essa estrutura não representa explicitamente o autor responsável por cada opinião. Em contextos com múltiplos indivíduos, essa ausência pode gerar ambiguidades na associação entre opiniões e seus respectivos autores [Wiebe et al. 2005]. A estrutura de anotação disponível para ASBA é ilustrada na Figura 1.

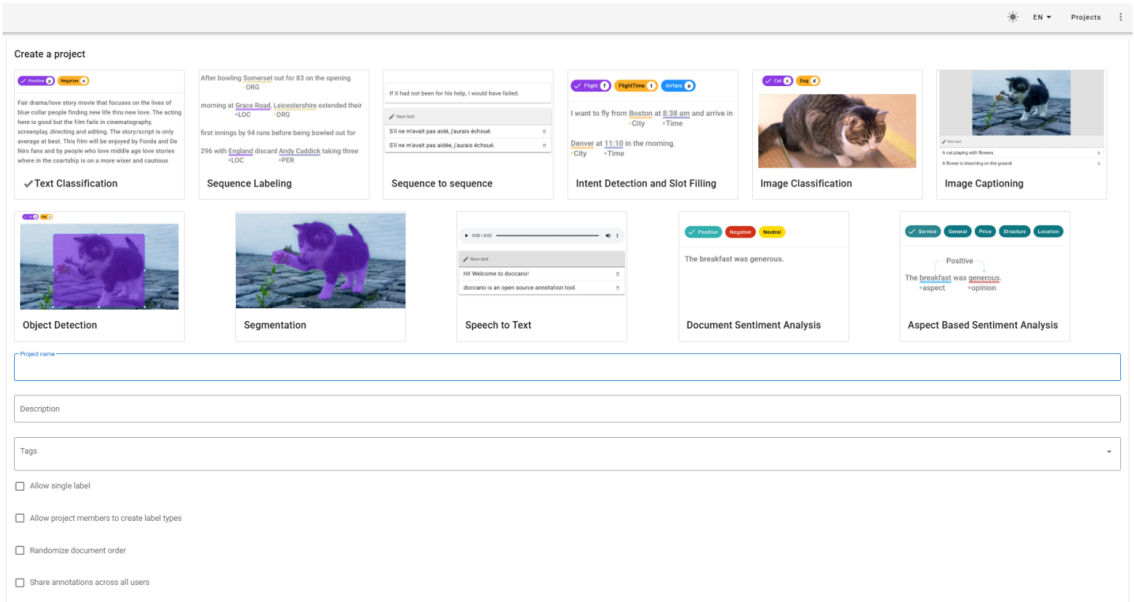


Figura 1. Anotação na ferramenta Doccano para a tarefa de ASBA.

2.2. Implementação

A implementação foi realizada sobre a versão do Doccano previamente expandida para ASBA [Cardozo 2025]. A extensão proposta incorpora o autor da opinião (*opinion holder*) à estrutura já existente, permitindo representar anotações compostas por aspecto, opinião, polaridade e autor.

A decisão de incorporar o autor como uma referência opcional associada à relação existente teve como objetivo preservar a estrutura já utilizada pela versão para ASBA, evitando a criação de um fluxo de anotação paralelo. Dessa forma, a extensão mantém a associação entre aspecto, opinião e polaridade e adiciona a autoria como informação complementar da mesma estrutura relacional.

Tabela 1. Principais alterações realizadas na extensão do Doccano.

Componente	Alteração
Modelo de dados	Inclusão do autor nas relações e persistência no banco de dados.
Importação	Leitura do autor e de sua posição no texto.
Exportação	Inclusão do autor nas anotações exportadas.
Frontend	Criação, associação e visualização da nova relação.

No *backend*, o modelo de relações foi estendido com uma referência opcional ao *span* que representa o autor, incluindo sua persistência no banco de dados. Os fluxos de importação, serialização e exportação também foram adaptados para processar essa informação. Em particular, a importação passou a reconhecer o autor e suas posições no texto, criar os tipos de *span* e relação necessários e persistir as anotações estruturadas. Também foi incluído um teste automatizado para verificar a preservação da informação de autoria no processo de exportação das anotações.

No *frontend*, a representação das relações e o fluxo de anotação foram modificados para incorporar o autor. Quando o anotador relaciona um autor a um aspecto, a interface não cria uma relação independente; em vez disso, atualiza a informação de autoria da relação aspecto–opinião correspondente e utiliza uma ligação auxiliar para representar visualmente essa associação. Essa estratégia preserva o fluxo original de anotação e evita a duplicação de relações no modelo de dados, conforme ilustrado na Figura 2.



(a) ASBA **(b) Extensão proposta**

Figura 2. Comparação entre a anotação ASBA (a) e a extensão proposta com representação explícita do autor da opinião (b).

2.3. Avaliação Preliminar

A avaliação preliminar teve como objetivo verificar se a extensão proposta permite representar explicitamente informações de autoria em situações nas quais a estrutura utilizada na versão para ASBA não oferece esse recurso. Para isso, foi realizado um estudo de caso com 30 avaliações de hotéis extraídas de um conjunto de dados previamente utilizado em tarefas de análise de sentimentos estruturada [Campos et al. 2026].

Os mesmos exemplos foram anotados por um único anotador nas duas versões da ferramenta. Em cada caso, foram mantidos os mesmos elementos de aspecto,

opinião e polaridade, permitindo observar especificamente a diferença introduzida pela representação do autor da opinião. Considerou-se como autor uma menção explicitamente identificável no texto à qual pudesse ser atribuída a opinião anotada.

Por exemplo, em um caso no qual uma pessoa é mencionada explicitamente como responsável por determinada opinião, a versão para ASBA representa apenas aspecto, opinião e polaridade, enquanto a extensão também registra a entidade responsável e sua associação à opinião.

Das 30 avaliações analisadas, 10 continham referência explícita a um autor da opinião. Nesses casos, apenas a extensão proposta permitiu representar e armazenar explicitamente a associação entre autor e opinião. O estudo permitiu verificar o funcionamento do novo esquema de anotação, embora uma avaliação com múltiplos anotadores ainda seja necessária para investigar aspectos como usabilidade e concordância interanotador.

3. Conclusão

Este trabalho apresentou uma extensão da versão do Doccano voltada à ASBA, incorporando a representação explícita do autor da opinião. A proposta adiciona esse elemento ao modelo de relações e adapta os fluxos de importação, exportação e anotação da ferramenta, permitindo representar aspecto, opinião, polaridade e o respectivo autor. A avaliação preliminar com avaliações de hotéis permitiu verificar o funcionamento dessa estrutura em comparação à versão utilizada como base. Das 30 avaliações analisadas, 10 continham um autor explicitamente identificável, casos em que a extensão permitiu armazenar a associação entre autor e opinião.

Como limitação, a avaliação foi conduzida por um único anotador, não sendo possível concluir sobre a usabilidade da ferramenta ou a concordância entre diferentes usuários. Como trabalhos futuros, pretende-se realizar experimentos com múltiplos anotadores, avaliar a concordância interanotador e investigar o impacto da informação de autoria em tarefas posteriores de análise de sentimento.

Referências

Barnes, J., Kurtz, R., Oepen, S., Øvrelid, L., and Velldal, E. (2021). Structured sentiment analysis as dependency graph parsing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3387–3402.

Campos, A. B., Corrêa, U. B., and Freitas, L. A. (2026). Structured sentiment analysis in brazilian portuguese: An exploratory study using bertimbau. In *Proceedings of the International Conference on Computational Processing of Portuguese (PROPOR)*. In Portuguese.

Cardozo, A. d. S. (2025). doccano: Ferramenta de anotação para análise de sentimentos expandida. `https://github.com/aleciocardozo/doccano`. Repositório GitHub.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., and Liang, X. (2018). doccano: Text annotation tool for human.

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Al-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., et al. (2016). Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 19–30.

Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39:165–210.

A Comparative Study of Discourse Signal Proposals in Rhetorical Structure Theory

Gabriela Fortalesa[1], Jackson Wilke da Cruz Souza[2]

[1] Departamento de Letras e Artes - Universidade Estadual de Feira de Santana
Feira de Santana, Bahia / Brasil

[2]Programa de Pós-Graduação em Língua e Cultura - Universidade Federal da Bahia
Salvador, Bahia / Brasil

{gabrielafortalesa,jackcruzsouza}@gmail.com

Abstract. *This paper presents a comparative literature review on the Discourse Signals (DSs) of Rhetorical Structure Theory (RST) across six languages. The analysis demonstrates that coherence signaling is not universal, depending heavily on linguistic typology and textual modality. It becomes evident that typologically distant languages and new modalities require profound structural, syntactic, and hypertextual adaptations of the theory's original taxonomy.*

1. Introduction

Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) has established itself as one of the most robust theoretical models for mapping textual organization and coherence. Furthermore, the theory has made significant contributions to a wide range of applications in the field of Natural Language Processing (NLP), especially in discourse processing (Souza, Cardoso and Rodrigues, 2024).

Historically, the identification of rhetorical relations and discourse signaling focused almost exclusively on explicit Discourse Markers (DMs), such as conjunctions and prepositions. This paradigm, however, was transformed by the foundational work of Das and Taboada (2018), who expanded the notion of rhetorical relation signaling by showing that coherence extends beyond traditional connectives, as illustrated in (1).

(1) John is tall. Mary is short.

In this example, the two sentences are implicitly related according to the traditional view, since there is no DM connecting them. However, it can be observed that the two propositions may be linked through the *Contrast* relation despite the absence of an explicit connection between them. In this case, more recent studies in the literature indicate that the signaling may occur through the antonymy between the lexical items “tall” and “short”. As a result, Das and Taboada (2018) formalized a comprehensive taxonomy of non-connective (DSs) encompassing morphological, syntactic, semantic, graphical, and genre-related features.

As pointed out by Liu and Zeldes (2023), RST trees and relations vary widely in the ways they are signaled. According to the authors, signaling depends heavily on text type and genre, as well as on cross-linguistic differences. However, it can be observed that the

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 556–561
October 19-22, 2026 ©2026 Association for Computational Linguistics

taxonomic proposals for DSs presented in previous studies often overlap and/or differ from one another, sometimes only with regard to the label assigned to a given type of signal. In this context, a cross-sectional investigation of the existing studies could enrich current taxonomic proposals, especially with respect to Portuguese.

With this in mind, the present study aims to examine how discourse signals, across different languages, may be adapted to distinct genre and text-type contexts within the scope of RST. To this end, this paper is organized into four sections in addition to this Introduction. Section 2 presents the methodology adopted in this bibliographic study. Section 3 discusses the results and findings. Finally, Section 4 presents the concluding remarks and outlines directions for future research based on this study.

2. Methodology

This study was primarily based on a bibliographic and bibliometric review methodology (Soares, Picolli, and Casagrande, 2018), aiming to conduct a comparative literature review. The bibliographic survey was carried out using the Google Scholar and ACL Anthology databases. To this end, combinations of keywords (including their variations and translations) such as "Rhetorical Structure Theory," "RST," "discourse signals," "signalling," and "discourse markers" were used.

The preliminary screening was conducted through the reading of titles and abstracts, as suggested by Moreira, Guimarães, and Tsunoda (2020). Based on this process, studies presenting clear evidence of the formulation, adaptation, or adoption of discourse signal proposals were selected. Additionally, the cross-reference tracking technique (snowballing) (Ecar *et al.*, 2020) was employed, whereby references to other texts served as indicators for identifying further relevant studies. As a result, a corpus of 47 studies was compiled.

For the construction of the final comparative framework, only articles that adopted RST as their exclusive theoretical framework were retained. In this sense, studies that, although based on RST, operated through theoretical hybridization and/or multi-layer annotation were likewise excluded from the comparative scope, leaving a final total of 11 papers.

3. Results and discussion

Following the bibliographic review, the studies included relied on a wide variety of corpus genres. For English, Das and Taboada (2018) compiled a journalistic corpus, whereas Gessler, Liu, and Zeldes (2019) adopted a multi-genre corpus comprising social media posts, tutorials, news articles, and interviews. Regarding Brazilian Portuguese, Pardo (2005) analyzed a corpus of academic (scientific) texts, Antonio (2017) used spoken language transcripts, the study by Rodrigues, Souza and Cardoso (2023) relied on journalistic texts, and Gama and Souza (2024) employed data collected from social media. For Spanish and Basque, Da Cunha et al. (2011, 2012) built a multi-genre corpus including scientific, academic, and didactic texts, as well as news reports. In Chinese, Cao and Iruskieta (2018) compiled a multi-genre corpus consisting of academic texts, news articles, and advertisements, whereas Peng, Liu, and Zeldes (2022) expanded this approach by

incorporating academic texts, interviews, news articles, and tutorials. Seemann and Scheffler (2025) adopted an experimental corpus composed of constructed items and sentences for German, while Toldova *et al.* (2017) employed a multi-genre Russian corpus containing magazine, journalistic, and blog texts.

STUDIES	EXAMPLE
Das and Taboada (2018) - English	**"Although** L&T hadn't raised money for 38 year"
Gessler, Liu and Zeldes (2019) - English	"Three issues begin trading **today** [...] one began **last week**"
Pardo (2005) - Brazilian Portuguese	"Software Engineering is an evolving discipline [...] **with the primary objective** of producing high-quality, low-cost software."
Antonio (2017) - Brazilian Portuguese	(You defeat the Indians, keep the land for yourself)
Rodrigues, Souza and Cardoso. (2023) - Brazilian Portuguese	Defense Minister Nelson Jobim **announced** (...) **that** economist Solange Vieira…
Gama and Souza (2024) - Brazilian Portuguese	Let's see if even on a weak day, it gains some strength... **RT** @Live_Trade: #ecor3 closing 15' above 12.53 is already a good sign
Da Cunha *et al.* (2011; 2012) - Spanish	"[Los niños aprenden matemáticas] [**al** resolver problemas.]" (Children learn mathematics **by** solving problems)
Cao and Iruskieta (2018) - Chinese	**Jǐnguǎn** hái méiyǒu qǔdé zuìzhōng jiéguǒ, **dànshì** wǒmen rènwéi gāi móxíng"(**Although** the final results have not yet been obtained, we believe that the model…)
Seemann and Scheffler (2025) - German	"**Nachdem** er **ja** immer gesagt hatte, ich konnte ihn jederzeit anrufen, habe ich das gestern auch gemacht." ('**Because** he always said I could call him anytime, I did just that yesterday ')
Toldova *et al.* (2017) - Russian	There is no example of a complete sentence in the article, only marker-relation pairs, such as "**однако/а/но**" ('however/but'), "который" ('which'); "поэтому" ('so/therefore')

Table 1. Some illustrative examples drawn from the studies.

From Table 1, all the examples were translated to English from original studies and or idioms. For Brazilian Portuguese (BP), the pioneering study by Pardo (2005) focused on the extraction of explicit DMs from academic texts in order to create the first automatic parser for the language. This was later expanded by Rodrigues, Souza and Cardoso (2023), who systematized the use of combined signals in journalistic texts. As a result, the most recent study focused on the mapping of implicit relations, local syntax, and adaptation to the digital environment by identifying cues such as markers combined with verbal tenses. Still regarding BP, the study by Antonio (2017) illustrates that, when moving away from the standard of formal written texts, the signaling proposal needed to encompass interactional phenomena in spoken language (such as intonation and pauses) and to initiate a hypertextual paradigm for

User-Generated Content (UGC). Concerning the study by Gama and Souza (2024) on social media, for instance, it is noteworthy that elements inherent to the digital platform (such as links, emojis, and retweets) function as independent and domain-specific DSs.

With regard to Russian, the same principle of mapping non-connective items was adopted. This decision guided the formulation of the lexical inventory proposed by Toldova *et al.* (2017) for the structuring of discourse trees. In that study, non-connective items were mapped based on their co-occurrence, resulting in an inventory of content words (lemmas). As for the studies on English, Gessler, Liu, and Zeldes (2019) expanded the original model in order to anchor signals to the exact tokens in the text, including, for example, synonymy and visual layout features. Studies on German indicate that the disambiguation of rhetorical relations frequently occurs through modal aspects in colloquial sentences, making use of strong particles such as ja and doch, for instance.

With regard to languages that are typologically distant from one another, the studies presented more substantial innovations based on the proposal by Das and Taboada (2018), as observed in the work on Chinese. The authors highlighted the requirement for paired connectives and the high frequency of implicit relations, evidenced by simultaneous double markers (such as *jinguan* and *danshi*) and prenominal relative clauses.

Although most causal and concessive relations are marked by explicit lexical connectives across all the languages analyzed, the survey reveals that discourse signaling is far from uniform. First, there is typological variation in the amount of marking required for the same relation: Spanish and Portuguese rely on a single concessive connective, whereas Chinese requires an obligatory correlative pair. Second, not every signal is lexical: in spoken BP, clause order alone is sufficient to mark Condition; and in Chinese, the syntactic position of a relative clause itself triggers the same-unit relation. Third, DSs, such as modal particles in German and platform-specific abbreviations ("RT") in BP, show that the category of DS extends beyond the closed class of traditional connectives. Finally, marker ambiguity (Spanish "*al*", Portuguese "*que*", German "*doch*") recurs across typologically distant languages, suggesting that the functional polysemy of DSs is a general phenomenon rather than a language-specific one.

4. Final remarks

This study presented a bibliographic review and a comparative mapping of how discourse signals have been categorized and adapted across different languages and text genres from a strictly RST-based perspective. The aim was to demonstrate that coherence transcends the use of traditional markers, requiring expanded signaling frameworks capable of accounting for structures and phenomena related to syntax and to features associated with different modalities, such as spoken language and digital media.

It should be emphasized that the present study falls within the scope of research dedicated to discourse processing, particularly in BP. Recent findings demonstrate the relevance of RST as a theoretical framework, covering a wide range of languages and text genres, while also enabling notable applications in NLP.

As future work, we suggest expanding the scope of analysis to include hybrid or multi-layer theoretical models of discourse processing beyond RST. This approach will enable the exploration of larger corpora and additional language families, as well as deeper contrastive analyses of discourse signals. By explicitly comparing different taxonomies, future research can assess the impact of their incorporation or adaptation in studies on BP. Finally, as this study did not aim to propose a taxonomy for DSs in BP, future work following our exploratory and comparative analyses will focus on developing a data-driven taxonomy for BP.

Acknowledgements

The project was supported by the Ministry of Science, Technology and Innovation, with resources of Law N. 8,248, of October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published as Residence in TIC 13, DOU 01245.010222/2022-44. In addition, this work received financial support from the FAPESB/CNPq Call for Proposals 004/2023 - First Projects Program.

References

Antonio, J. D. (2017). Mecanismos utilizados pelos destinatários do discurso para identificação de relações de coerência não sinalizadas por conectores. *DELTA: Documentação de Estudos em Linguística Teórica e Aplicada*, 33(1), p. 79-108. DOI: 10.1590/0102-445025798334674077

Cao, S., da Cunha, I., & Iruskieta, M. (2018). The RST Spanish-Chinese treebank. In: *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, p. 156-166.

Da Cunha, I., Torres-Moreno, J. M., & Sierra, G. (2011). On the development of the RST Spanish Treebank. In: *Proceedings of the 5th Linguistic Annotation Workshop*, p. 1-10.

Da Cunha, I., et al. (2012). DiSeg 1.0: The first system for Spanish discourse segmentation. *Expert Systems with Applications*, 39(2), p. 1671-1678. DOI: 10.1016/j.eswa.2011.06.058

Dantas, E., Santa Bárbara, L., Pereira, M., Gama, N., Almeida, T., Souza, J. W. C., Cardoso, P. C. F., & Rodrigues, R. (2024). *Manual de anotação de sinalizadores discursivos em textos jornalísticos*. Relatório Técnico nº 447, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

Das, D., & Taboada, M. (2018). RST Signalling *Corpus*: A corpus of signals of coherence relations. *Language Resources and Evaluation*, 52(1), p. 149-184. DOI: 10.1007/s10579-017-9383-x

Ecar, M. *et al.* (2020) Software process improvement diagnostic: A snowballing systematic literature review. *In: 2020 XLVI Latin American Computing Conference (CLEI)*. IEEE. p. 156-164. DOI: 10.1109/CLEI52000.2020.00025

Gama, N. S., & Souza, J. W. C. (2024). *Estudo preliminar sobre sinalizadores discursivos para Conteúdo Gerado por Usuários*. Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana. DOI: 10.5753/stil.2024.245060.

Gessler, L., Liu, Y. J., & Zeldes, A. (2019). A discourse signal annotation system for RST trees. In: *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, p. 56-61. DOI: 10.18653/v1/W19-2708

Liu, Y. J., & Zeldes, A. (2023). Why Can't Discourse Parsing Generalize? A Thorough Investigation of the Impact of Data Diversity. In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, p. 3112-3130. DOI: 10.48550/arXiv.2302.06488

Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3), p. 243-281.

Moreira , P.S.C.; Guimarães, A. J. R. & Tsunoda, D. F. (2020). Qual ferramenta bibliométrica escolher? Um estudo comparativo entre softwares. P2P e Inovação, [s.l.], v. 6, p. 140-158. DOI: 10.21721/p2p.2020v6n2.p140-158

Pardo, T. A. S. (2005). *Métodos para análise discursiva automática*. Tese (Doutorado em Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. DOI: 10.11606/T.55.2005.tde-29082005-172336

Peng, S., Liu, Y. J., & Zeldes, A. (2022). GCDT: A Chinese RST Treebank for Multigenre and Multilingual Discourse Parsing. In: *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, p. 382-391. DOI: 10.18653/v1/2022.aacl-short.47

Rodrigues, R., Souza, J. W. C., & Cardoso, P. C. F. (2023). Sinalizadores retórico-discursivos: revisitando a anotação RST no córpus CSTNews. In: *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, SBC, p. 249-257. DOI: 10.5753/stil.2023.234120

Seemann, H. J., & Scheffler, T. (2025). German modal particles as discourse Signals. *Dialogue & Discourse*, v. 16, n. 1, p. 1-30. DOI: 10.5210/dad.2025.101

Soares, S. V., Picolli, I. R. A. & Casagrande, J. L (2018). Pesquisa bibliográfica, pesquisa bibliométrica, artigo de revisão e ensaio teórico em administração e contabilidade. *Administração: ensino e pesquisa*, v. 19, n. 2, p. 308-339. DOI: 10.13058/raep.2018.v19n2.970

Souza, J. W. C., Cardoso, P. C. F., & Rodrigues, R. (2024). Systematic Review of Studies on Rhetorical Structure Theory. *Revista de Estudos da Linguagem*, 31(3), p. 1643–1675. DOI: 10.17851/2237-2083.31.3.1643-1675

Toldova, S., *et al.* (2017). Rhetorical relation markers in Russian RST Treebank. In: *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms*, p. 29-33. DOI: 10.18653/v1/W17-3604

Identificação e classificação de trechos racistas e sexistas em processos judiciais

Yasmin Francisquetti Barnes[1], **Marcelo Finger**[1]

[1]Instituto de Matemática e Estatística – Universidade de São Paulo (USP)
05508-090 – São Paulo – SP – Brasil

yasminfrancisquetti@usp.br, mfinger@ime.usp.br

Abstract. *This study describes the development of an artificial intelligence tool to identify and classify racist and sexist discourses from legal agents. Focusing on sexual violence lawsuits at the São Paulo State Court of Justice (TJSP), the research aims to highlight biases that corroborate secondary victimization. The computational solution relies on a Natural Language Processing model based on the Transformer architecture, utilizing a BERT-derived model pre-trained for Brazilian Portuguese. The model is trained with textual data categorized by a gender stereotype ontology, and its performance is evaluated through 5-fold cross-validation using accuracy, precision, recall, and F-measure.*

Resumo. *Este estudo descreve o desenvolvimento de uma ferramenta de inteligência artificial para identificar e classificar discursos racistas e sexistas proferidos por agentes jurídicos. Focado em processos de violência sexual do Tribunal de Justiça de São Paulo (TJSP), a pesquisa visa evidenciar vieses que corroboram a vitimização secundária. A solução computacional baseia-se em um modelo de Processamento de Linguagem Natural com arquitetura Transformer, utilizando um modelo derivado do BERT pré-treinado para o português brasileiro. O modelo é treinado com dados textuais categorizados por uma ontologia de estereótipos de gênero e seu desempenho é avaliado por validação cruzada ($k = 5$) com métricas de acurácia, precisão, revocação e medida-F.*

1. Introdução

Em 2024, foram registrados 83.114 casos de estupro no Brasil, representando o maior número de ocorrências entre os 5 últimos anos [Ministério da Justiça e Segurança Pública 2025] e, dentre esses casos, 86% das vítimas são do sexo feminino. Entretanto, historicamente, o estupro em si raramente constitui uma violação isolada: após o crime, a vítima normalmente é submetida a diversas manifestações de violência simbólica, tanto institucionais quanto sociais [Siqueira 2016, Sousa 2017].

No âmbito judicial, o momento após o ataque é crucial para a manifestação da vítima e o estabelecimento do processo jurídico que tratará o caso. No entanto, é justamente durante a busca pela justiça que um novo ciclo de violência se inicia, uma vez que o Sistema de Justiça Criminal está condenado, em suas raízes, a reproduzir comportamentos derivados do patriarcalismo presente na sociedade [Siqueira 2016]. Essa experiência, caracterizada pelo tratamento inadequado, insensível ou deslegitimador por

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 562–566
October 19-22, 2026 ©2026 Association for Computational Linguistics

parte das instituições judiciais, é conhecida pela Criminologia como vitimização secundária [Conselho Nacional do Ministério Público sd]. Nela, a denúncia da vítima e suas reivindicações são tratadas com negligência e de maneira desumana.

Nesse contexto, a análise manual do elevado volume de processos em andamento para identificar padrões discursivos potencialmente enviesados torna-se inviável, demandando soluções computacionais capazes de automatizar a triagem de dados textuais em larga escala. Para esse fim, o Processamento de Linguagem Natural (PLN) oferece ferramentas consolidadas para a detecção de discursos nocivos em português brasileiro, incluindo linguagem ofensiva, misoginia e discurso de ódio. Trabalhos de referência nessa tarefa, como os de Pelle e Moreira [Pelle e Moreira 2017], Leite et al. [Leite et al. 2020] e Vargas et al. [Vargas et al. 2022], empregam desde modelos clássicos até arquiteturas baseadas em Transformers, como o BERTimbau. A aplicação dessas técnicas ao domínio jurídico viabiliza o mapeamento sistemático de padrões discursivos em grandes bases processuais, fornecendo subsídios para políticas públicas e práticas formativas.

Tendo isso em vista, o presente estudo busca a elaboração de um modelo de rede neural que auxiliará na classificação dos discursos proferidos por agentes do direito em processos de violência sexual, identificando discurso nocivo e desumano nos âmbitos de gênero e raça.

2. Metodologia

Este estudo apresenta o desenvolvimento de um modelo de processamento de linguagem natural, baseado em rede neural, treinado a partir de um *corpus* composto por processos judiciais de casos de estupro em formato textual, contendo trechos anotados com manifestações de viés. Os processos que formam esse *corpus* estão sob segredo de justiça do Tribunal de Justiça do Estado de São Paulo (TJSP) e seu acesso foi autorizado para a pesquisa, mediante a assinatura de Termo de Confidencialidade e Sigilo por todos os pesquisadores e colaboradores envolvidos no projeto. O processo de anotação é realizado por uma equipe interdisciplinar, que identifica trechos de acordo com categoria, conotação e alvo pré-definidos, embasados e adaptados da ontologia estruturada por Marques [Marques 2021].

A categorização dos trechos identificados como portadores de estereótipo é estruturada em eixos principais e subcategorias. Os itens avaliados englobam: o conceito do crime de estupro (incluindo noções muitas vezes equivocadas de consentimento, menosprezo ao crime, violência e reação da vítima); discussões sobre provas; conduta pré-fato (abrangendo conduta sexual e vestimenta); e a vida pregressa (englobando contexto familiar, profissão e sexualidade). Além disso, mapeiam-se a descredibilização da vítima (por meio de alegadas inconsistências no depoimento ou falsa denúncia por vingança), o estado mental, a vulnerabilidade (envolvendo entorpecimento/embriaguez e grau de consciência) e a responsabilidade parental, que em geral ocorre atribuindo-se à mãe uma negligência pela ocorrência da violência.

Após a categorização, estabelece-se a conotação do trecho, dividida entre qualificadora (fala de cunho positivo com base na categorização padrão da sociedade) ou desqualificadora (cunho negativo). Por fim, define-se o alvo da fala, que engloba as possibilidades de vítima, pessoa agressora, mulher, homem, pessoa genérica, ou um conceito/ideia. Em paralelo com a anotação, realiza-se uma contínua análise dessas rotulações para a

melhor definição dos parâmetros de entrada e saída do modelo.

O *corpus* em construção é composto por cerca de 300 processos do TJSP. Atualmente, a base encontra-se em uma etapa de aprimoramento, passando por uma nova extração automatizada via portal e-SAJ com o intuito de maximizar a captura de informações. No estágio atual da pesquisa, a etapa de anotação encontra-se em fase piloto, contando com quatro processos integralmente rotulados por pesquisadoras especialistas da área do Direito e três em fase de processamento. Cada trecho é avaliado de maneira independente por, em média, dois anotadores (raramente três). Devido ao volume baixíssimo de trechos anotados nesta etapa preliminar, e considerando que a ontologia definitiva ainda será aplicada sistematicamente aos novos processos extraídos, a apresentação da distribuição estatística das categorias e a mensuração da concordância interanotadores serão realizadas apenas após a consolidação da base. Pretende-se aplicar o Coeficiente Kappa de Cohen [Cohen 1960] caso o protocolo de anotação seja consolidado com dois anotadores, ou uma métrica equivalente caso haja variação no *corpus*.

Em relação à arquitetura, estabelece-se um modelo de rede neural *Transformer*, baseado em mecanismos de atenção (*self-attention*). Em específico, implementa-se um modelo da família BERT [Devlin et al. 2019]. Visando o processamento de textos jurídicos em português brasileiro, adota-se inicialmente o modelo pré-treinado BERTimbau [Souza et al. 2020] como *baseline*. Adicionalmente, se viável no decorrer da pesquisa, será realizada uma avaliação comparativa de desempenho com outros modelos recentes para o português, como o ModernBERT-pt-BR [Wu e Garcia 2025] e o NorBERTo [Pellicer e Rinaldo 2024]. A partir dessa base, desenvolve-se o classificador automático.

Para a aplicação do modelo, adota-se o método de validação k-fold *cross-validation* com $k = 5$. O modelo passa por 5 turnos, sendo treinado com 80% do *corpus* disponível e testado com os 20% restantes, mantendo a separação o mais balanceada possível entre as categorias. Para avaliar a pertinência da rede, seu funcionamento é medido globalmente e por categoria através das métricas de acurácia, precisão, revocação e medida-F. Para os fins da pesquisa, classifica-se como "positivo" o evento no qual realmente existe uma expressão nociva no discurso, e "negativo" como um discurso neutro.

A estrutura geral da metodologia proposta, abrangendo desde a etapa de coleta de dados até a avaliação das métricas do modelo, é apresentada na Figura 1.

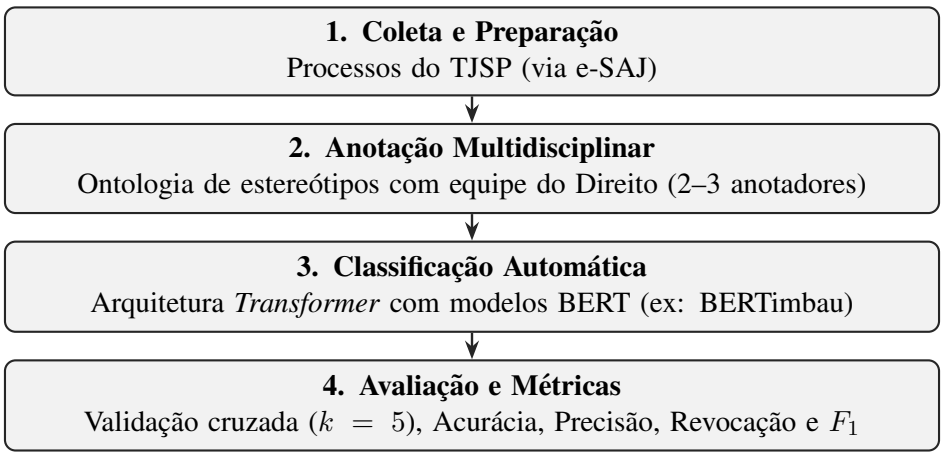


Figura 1. Fluxo metodológico geral da solução computacional proposta.

3. Resultados Esperados

Como resultado da investigação de possíveis manifestações de racismo e sexismo em discursos proferidos por agentes do sistema de justiça, esperam-se impactos de relevância tanto científica quanto social. Do ponto de vista social, a identificação desses padrões contribuirá para a compreensão empírica dos mecanismos institucionais associados à vitimização secundária em casos de violência sexual.

Do ponto de vista científico, espera-se contribuir para o desenvolvimento de métodos computacionais aplicados à análise de linguagem no domínio jurídico, particularmente no contexto do processamento de linguagem natural voltado à identificação de discurso nocivo. Ferramentas dessa natureza auxiliarão pesquisadores e instituições na análise de grandes volumes de documentos judiciais, permitindo investigações que seriam inviáveis por meio de análise manual.

Dessa forma, a consolidação do modelo de rede neural para a classificação de discursos racistas e sexistas em processos judiciais fornecerá um instrumento que auxilie na análise sistemática desses documentos, contribuindo efetivamente para estudos sobre viés institucional e para o aprimoramento de práticas operacionais no sistema de justiça.

4. Considerações Finais

Em suma, este estudo propõe uma abordagem interdisciplinar para a identificação de vieses racistas e sexistas no judiciário brasileiro, integrando técnicas de Processamento de Linguagem Natural à análise do discurso jurídico. A estruturação de um classificador automático baseado na arquitetura *Transformer*, alimentado por processos de crimes de violência sexual do TJSP, representa um avanço prático para a detecção em larga escala de falas associadas à vitimização secundária.

Ainda que a lida com dados sensíveis e sob segredo de justiça imponha desafios inerentes de acesso e processamento, a metodologia de anotação ancorada em ontologias de estereótipos confere rigor à ferramenta em desenvolvimento. Ao aplicar o aprendizado de máquina à revisão processual, a pesquisa oferece um método quantitativo e estruturado para observar a linguagem jurídica. Dessa forma, o projeto viabiliza a extração de métricas concretas sobre a presença de vieses nos autos, fornecendo recursos computacionais que facilitam a análise empírica desses documentos.

Referências

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Conselho Nacional do Ministério Público (s.d.). Vitimização. `https://www.cnmp.mp.br/defesadasvitimas/vitimas/vitimizacao`. Acesso em: 3 ago. 2026.

Devlin, J., Chang, M., Lee, K., e Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. Em *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, pages 4171–4186. Association for Computational Linguistics.

Leite, J. A., Silva, D., Bontcheva, K., e Scarton, C. (2020). Toxic language detection in social media for Brazilian Portuguese: Task and dataset with multilabel annotation. Em *Proceedings of the 4th Workshop on Online Abuse and Harms (ALW)*, pages 1–13.

Marques, C. S. (2021). Uma ontologia para permitir a análise da presença e influência do estereótipo de gênero em processos jurídicos de casos de estupro. Trabalho de conclusão de curso (bacharelado em ciência da computação), Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo.

Ministério da Justiça e Segurança Pública (2025). Mapa de segurança pública 2025. Relatório técnico, Ministério da Justiça e Segurança Pública, Brasília, DF.

Pelle, R. d. e Moreira, V. (2017). Offensive comments in the Brazilian web: a dataset and baseline results. Em *Anais do VI Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*, pages 510–519, Porto Alegre, RS, Brasil. SBC.

Pellicer, L. F. A. O. e Rinaldo, G. (2024). NorBERTo: A ModernBERT model trained for Portuguese with 331 billion tokens corpus. Em *Proceedings of the International Conference on the Computational Processing of Portuguese (PROPOR)*.

Siqueira, C. K. B. (2016). A liberdade sexual da mulher na prática judicial: análise da aplicação de estereótipos de gênero em processos de estupro. Dissertação de mestrado, Faculdade de Direito, Universidade Federal do Ceará, Fortaleza.

Sousa, R. F. d. (2017). Cultura do estupro: prática e incitação à violência sexual contra mulheres. *Estudos Feministas*, 25(1):9–29.

Souza, F., Nogueira, R., e Lotufo, R. (2020). BERTimbau: Pretrained BERT models for Brazilian Portuguese. Em *Intelligent Systems: 9th Brazilian Conference (BRACIS)*, pages 403–417. Springer International Publishing.

Vargas, F., Carvalho, I., Rodrigues de Góes, F., Pardo, T., e Benevenuto, F. (2022). HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. Em Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., e Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7174–7183, Marseille, France. European Language Resources Association.

Wu, W. B. T. L. e Garcia, L. P. F. (2025). ModBERTBr: A ModernBERT-based model for Brazilian Portuguese. Em *Anais do Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*.

Data-Faithful Natural Language Generation for Explaining Fitness-Market Anomalies

Kauan Divino Pouso Mariano[1], **Fabrycio Leite Nakano Almada**[1]
Victor Emanuel da Silva Monteiro[1], **Maykon Adriell Dutra**[1]
Jefferson Felex de Faria[2], **Jennifer Vitória da Silva Peixoto**[2]
Kamylla Sejane Pouso Freitas[3]

[1]Institute of Informatics – Federal University of Goiás (UFG)
Goiânia – GO – Brazil

[2]Polytechnic School – Pontifical Catholic University of Goiás (PUC)
Goiânia – GO – Brazil

[3]School of Medicine – Alfredo Nasser University Center
Aparecida de Goiânia – GO – Brazil

{kauan,fabrycio,victor_emanuel,maykonadriell}@discente.ufg.br

felexmc22@gmail.com, jenniferpeixoto04@gmail.com

kamyllasejane@unifan.edu.br

Abstract. *This study presents an integrated pipeline that transforms anomaly-detection outputs from global fitness-market data into data-faithful narratives. Using a 132-country panel and a 2019 baseline, we analyze observed post-COVID-19 recovery through 2025, while treating 2026 only as a modelled extension. Four complementary detectors—regional-median deviations, linear-regression residuals, Random Forest residuals, and Isolation Forest—are combined through method agreement and mapped to a rule-based typology. A deterministic, template-based Natural Language Generation (NLG) component then produces short, medium, and complete explanations from tabular evidence. In 2025, 14 countries were consensus anomalies; Guyana exemplified market growth without participatory expansion, with revenue recovery of +401.43%, participation recovery of -2.06%, and a 403.50-point gap. All 132 narratives passed automatic factual-consistency checks. These checks validate field preservation rather than fluency, usefulness, or explanatory quality; human evaluation remains future work.*

1. Introduction

COVID-19 disrupted physical activity through restrictions on mobility, gym access, and daily routines [Hasson et al. 2022]. Recovery remained uneven across populations and contexts [Eime et al. 2024, Strain et al. 2024]; although the fitness sector adapted and resumed growth, its expansion did not necessarily yield proportional gains in population participation [Cecere and Bernardi 2023, Liu et al. 2022].

This heterogeneity creates an analytical challenge for global fitness-market data, which combine market, participation, socioeconomic, health, and recovery indicators.

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 567–572
October 19-22, 2026 ©2026 Association for Computational Linguistics

Rankings or descriptive statistics may hide relevant deviations, especially when economic recovery diverges from participation trends. This work investigates how anomaly-detection outputs from global fitness-market indicators can be transformed into scientifically grounded explanatory narratives. The proposed pipeline integrates tabular anomaly detection with NLG: anomaly detection identifies countries whose recovery profiles deviate from regional, contextual, or multivariate expectations, and controlled templates convert these signals into data-faithful narratives.

The approach is grounded in data-to-text generation, explainable analysis, and anomaly explanation. Data-to-text research emphasizes that NLG from structured data requires content selection, ordering, interpretation, and factual consistency evaluation, not only fluent verbalization [Lin et al. 2024]. As related work, LIDA provides a conceptual reference for multi-stage data-communication pipelines based on semantic summarization, goal identification, and faithful generation, but it was not used in our implementation [Dibia 2023]; data-storytelling research further highlights the narrative organization of data communication [Ren et al. 2023]. In parallel, Explainable Artificial Intelligence and outlier-explanation studies argue that anomalous observations need interpretable descriptions of the variables responsible for their deviation [Mersha et al. 2024, Tritscher et al. 2023, Panjei et al. 2022]. Rather than proposing a new anomaly-detection algorithm, the methodological contribution integrates four detectors, method-agreement consensus, a rule-based typology, controlled NLG, and factual-consistency checks to communicate post-COVID-19 divergences among market recovery, population participation, and health-related indicators.

2. Methodology

The analysis uses `data/clean_gym_data.csv`, a cleaned country-year panel derived from the Kaggle World Gym & Fitness Trends (2000–2026) dataset and covering 132 countries.[1] Its documentation lists the Health & Fitness Association, World Health Organization Global Health Observatory, World Bank Open Data, EuropeActive/Deloitte, and Our World in Data as institutional sources; however, this study did not reconstruct the primary collection or harmonization process. The panel combines market and socioeconomic or health indicators, later aggregated into country-level recovery features; global participation is summarized as a simple, unweighted mean across countries. The main analysis compares 2022–2025 with the 2019 pre-COVID-19 baseline, whereas 2026 is treated as a modelled extension rather than a consolidated observation. Cross-country comparisons may therefore reflect unresolved differences in provenance, harmonization, inflation adjustment, and currency conversion.

The pipeline follows five compact stages. First, revenue recovery, participation recovery, and `market_health_gap` are computed relative to 2019; the gap is defined as revenue recovery minus participation recovery, with positive values indicating stronger economic recovery than participation recovery and negative values indicating the opposite. This indicator is not treated as a direct public-health measure, but as an interpretable contrast linking anomaly detection to narrative explanation. Second, anomaly detection combines four complementary methods: regional-median deviations, linear-regression residuals, Random Forest residuals [Li and van Leeuwen 2023], and Isolation

[1] Implementation scripts are available at `https://github.com/kauandivino/globalfit-anomalies`.

Forest [Agyemang 2024]. Third, each method flags the top 10% of countries, with the number of cases rounded up. Method scores are z-standardized and averaged with equal weights into a combined score, while agreement is counted separately. A consensus anomaly is flagged by at least two methods; this operational rule reduces dependence on a single detector but is not a universal statistical threshold.

Fourth, a rule-based typology maps each deviation to an interpretable anomaly category and guides content selection. Finally, a deterministic, template-based NLG component produces short, medium, and complete narratives. Content realization is cumulative, as summarized in Table 1: each length preserves the preceding evidence and adds contextual slots populated directly from the anomaly and feature tables. The generator uses neither a Large Language Model nor free-form generation; predefined clauses and tabular slots constrain its output to explicit evidence. Automatic factual-consistency checks verify country and anomaly-type mentions, rounded numerical values, gap values, and absence of predefined causal terms, following data-to-text principles of content planning and factual consistency [Thomson et al. 2023, Alonso and Agirre 2024].

Table 1. Incremental content plan for the three deterministic NLG variants.

Variant	Required evidence slots	Interpretive and contextual content
Short	Country, anomaly type, revenue recovery, participation recovery, and market-participation gap	Core interpretation selected by the rule-based typology
Medium	All short-version slots	Region and deviations from the regional median and model-based expectation
Complete	All medium-version slots	Selected 2019 socioeconomic and health indicators, plus the model-based reference value

3. Results

Between 2019 and 2025, global health-club revenue increased by 69.97%, the number of gyms by 19.79%, and memberships by 26.21%, whereas the simple country mean of fitness participation rose only 0.38 percentage points, from 40.24% to 40.62%. This asymmetry shows that market recovery and population participation are related but not equivalent dimensions.

The anomaly-detection stage identified 14 consensus anomalies in 2025, four of which were flagged by all four methods. Residual-model validation was weak: participation obtained R^2 values of approximately 0.13–0.18, while revenue and `market_health_gap` had negative R^2 values; residuals are therefore treated as exploratory expected-behavior references rather than strong predictions. Sensitivity tests yielded 6, 14, and 19 consensus anomalies at 5%, 10%, and 15% thresholds, respectively. Guyana, the United States of America, India, and Lebanon remained consensus anomalies across all three settings, indicating a stable core but a threshold-dependent complete set. Figure 1 shows the 2025 relation between revenue and participation recovery, while Table 2 summarizes selected profiles.

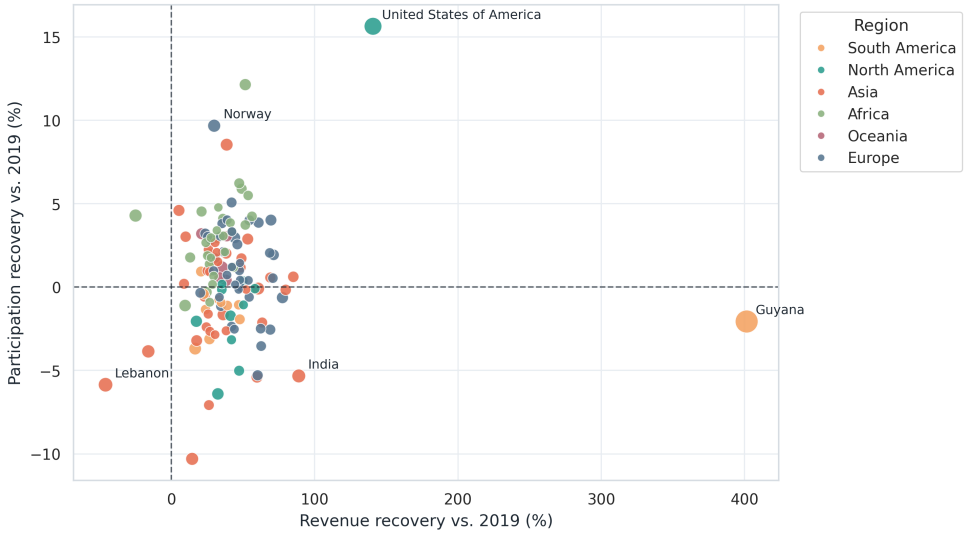


Figure 1. Revenue versus participation recovery by country in 2025. Colors denote regions, bubble size reflects the combined anomaly score, and labels identify selected consensus anomalies.

Table 2. Selected 2025 anomaly profiles; recovery is relative to 2019 and gap is in percentage points.

Country	Type	Agreement	Rev. (%)	Part. (%)	Gap (p.p.)
Guyana	Market growth without participatory expansion	4/4	+401.43	-2.06	403.50
Lebanon	Market recovery below regional pattern	4/4	-45.95	-5.85	-40.10
Norway	Participation above expectation with low revenue recovery	2/4	+29.85	+9.68	20.17

The agreement column reports how many of the four detectors flagged each country: Guyana and Lebanon represent unanimous 4/4 agreement, whereas Norway illustrates the minimum 2/4 operational consensus. The table also shows the sequence from evidence to explanation. For Guyana, revenue recovery of +401.43%, participation recovery of -2.06%, and a 403.50-point gap triggered the type *market growth without participatory expansion*. The corresponding corrected English template rendering is: *"Guyana showed market growth without participatory expansion in 2025: fitness-sector revenue increased by 401.43% relative to 2019, participation decreased by 2.06%, and the market-participation gap reached 403.50 percentage points. This profile represents strong sectoral expansion without proportional population engagement."* Thus, the rule-based label selects the interpretation, while tabular values provide its evidence. Because relative changes may be amplified by small initial values, the extreme percentage is treated as an anomaly signal rather than externally validated national evidence.

In a 2025 pipeline run, all 132 narratives passed the automatic factual-consistency checks, with a mean score of 1.0 for country and anomaly-type mentions, rounded recovery and gap values, and absence of predefined causal terms. These checks validate template filling and factual consistency, but do not assess fluency, clarity, usefulness, or

explanatory quality, and therefore do not replace human evaluation.

4. Discussion and Final Remarks

The applied contribution integrates complementary anomaly detectors, consensus, rule-based typology, deterministic NLG, and factual-consistency checks, making explanations traceable to explicit indicators. Fitness-market recovery should not be equated with public-health improvement [Duijvestijn et al. 2023]. However, the analysis identifies associations rather than causal effects; residual models are exploratory, the complete anomaly set is threshold-sensitive, and cross-country interpretation remains limited by unresolved provenance and harmonization issues.

Template constraints do not demonstrate lower hallucination rates than alternative generators, and a perfect factual-consistency score verifies selected inputs rather than overall textual quality. Future work should add human evaluation [Thomson et al. 2024], compare templates with Large Language Models and neural or free-form generation, assess unsupported content, externally validate national indicators, improve expected-value models, and test other domains.

Artificial Intelligence Usage

Generative artificial intelligence (AI) tools were used during manuscript preparation to support language revision, technical documentation, LaTeX editing, and presentation. Specifically, Prism, ChatGPT, and Claude, including reasoning-enabled modes, were used to paraphrase excerpts, improve English clarity, draft and revise LaTeX table structures, review textual consistency with reported outputs, and refine figure captions. These tools were editorial and technical aids only; they were not part of the implemented NLG pipeline and did not generate the narratives evaluated in this study. They did not determine the research questions, analytical methods, empirical results, or scientific interpretations. All AI-assisted content was reviewed and validated by the authors, who assume full responsibility for the manuscript.

References

Agyemang, E. F. (2024). Anomaly detection using unsupervised machine learning algorithms: A simulation study. *Scientific African*, 26:e02386.

Alonso, I. and Agirre, E. (2024). Automatic logical forms improve fidelity in table-to-text generation. *Expert Systems with Applications*, 238:121869.

Cecere, R. and Bernardi, P. (2023). From crisis to opportunity: exploring the interplay between entrepreneurial processes and technological changes in italian fitness industry. *International Journal of Entrepreneurship and Innovation Management*, 27(3/4):237–268.

Dibia, V. (2023). LIDA: A tool for automatic generation of grammar-agnostic visualizations and infographics using large language models. In Bollegala, D., Huang, R., and Ritter, A., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 113–126, Toronto, Canada. Association for Computational Linguistics.

Duijvestijn, M., de Wit, G. A., van Gils, P. F., and Wendel-Vos, G. C. W. (2023). Impact of physical activity on healthcare costs: a systematic review. *BMC Health Services Research*, 23(1):572.

Eime, R., Harvey, J., and Charity, M. (2024). Australian sport and physical activity behaviours pre, during and post-covid-19. *BMC Public Health*, 24(1):834.

Hasson, R., Sallis, J. F., Coleman, N., Kaushal, N., Nocera, V. G., and Keith, N. (2022). Covid-19: Implications for physical activity, health disparities, and health equity. *American Journal of Lifestyle Medicine*, 16(4):420–433.

Li, Z. and van Leeuwen, M. (2023). Explainable contextual anomaly detection using quantile regression forests. *Data Mining and Knowledge Discovery*, 37(6):2517–2563.

Lin, Y., Ruan, T., Liu, J., and Wang, H. (2024). A survey on neural data-to-text generation. *IEEE Transactions on Knowledge and Data Engineering*, 36(4):1431–1449.

Liu, R., Menhas, R., Dai, J., Saqib, Z. A., and Peng, X. (2022). Fitness apps, live streaming workout classes, and virtual reality fitness for physical activity during the covid-19 lockdown: An empirical study. *Frontiers in Public Health*, Volume 10 - 2022.

Mersha, M., Lam, K., Wood, J., AlShami, A. K., and Kalita, J. (2024). Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction. *Neurocomputing*, 599:128111.

Panjei, E., Gruenwald, L., Leal, E., Nguyen, C., and Silvia, S. (2022). A survey on outlier explanations. *The VLDB Journal*, 31(5):977–1008.

Ren, P., Wang, Y., and Zhao, F. (2023). Re-understanding of data storytelling tools from a narrative perspective. *Visual Intelligence*, 1(1):11.

Strain, T., Flaxman, S., Guthold, R., Semenova, E., Cowan, M., Riley, L. M., et al. (2024). National, regional, and global trends in insufficient physical activity among adults from 2000 to 2022: a pooled analysis of 507 population-based surveys with 5.7 million participants. *The Lancet Global Health*, 12(8):e1232–e1243.

Thomson, C., Reiter, E., and Belz, A. (2024). Common flaws in running human evaluation experiments in nlp. *Computational Linguistics*, 50(2):795–805.

Thomson, C., Reiter, E., and Sundararajan, B. (2023). Evaluating factual accuracy in complex data-to-text. *Computer Speech & Language*, 80:101482.

Tritscher, J., Krause, A., and Hotho, A. (2023). Feature relevance xai in anomaly detection: Reviewing approaches and challenges. *Frontiers in Artificial Intelligence*, Volume 6 - 2023.

OABench-Consumidor: Um benchmark para avaliação de modelos de linguagem em Direito do Consumidor

Lucas B. Bulcão Mota[1], **Aline Athaydes**[1], **Babacar Mane**[1], **Daniela Barreiro Claro**[1]

[1]FORMAS Research Center on Data and Natural Language
Institute of Computing – Federal University of Bahia (UFBA)
Av. Milton Santos, s/n - Campus de Ondina – 40.170-110 – Salvador – BA – Brazil

{lucasbulcao,alineathaydes,babacarm,dclaro}@ufba.br

Abstract. *Evaluating language models in specific domains, such as Consumer Law, is challenging due to the lack of standardized resources in Brazilian Portuguese. To address this, we introduce OABench-Consumidor, an evaluation benchmark composed exclusively of Consumer Law questions from the Unified Bar Examination (OAB) covering editions from 2010.1 to 2025.2. This paper describes the data collection and structuring process. Furthermore, we demonstrate the benchmark's utility by evaluating the Qwen3-8B foundation model and a fine-tuned version. The specialized model achieved 64.84% accuracy (59/91), outperforming the baseline's 51.65% (47/91). The dataset provides a specialized resource for assessing Artificial Intelligence systems in this legal subarea.*

Resumo. *Avaliar modelos de linguagem em domínios específicos, como o Direito do Consumidor, é um desafio devido à escassez de recursos padronizados em português brasileiro. Para resolver esse problema, introduzimos o OABench-Consumidor, um benchmark composto exclusivamente por questões de Direito do Consumidor da primeira fase do Exame de Ordem Unificado (OAB), cobrindo as edições de 2010.1 a 2025.2. Este artigo descreve a coleta e a estruturação dos dados. Além disso, a utilidade do benchmark é demonstrada através da avaliação do modelo fundacional Qwen3-8B e de uma versão ajustada. O modelo especializado obteve 64,84% de acurácia (59/91), superando os 51,65% (47/91) do modelo base. O benchmark oferece um recurso direto para avaliar sistemas de Inteligência Artificial nessa subárea do Direito.*

1. Introdução

Nos últimos anos, os Grandes Modelos de Linguagem (LLMs) apresentaram avanços significativos em diversas tarefas de Processamento de Linguagem Natural (PLN). Apesar disso, avaliar o desempenho desses modelos no setor jurídico continua sendo um desafio. No Brasil, o desenvolvimento de sistemas de Inteligência Artificial enfrenta a falta de bases de dados e *benchmarks* em português.

O desenvolvimento de sistemas jurídicos confiáveis é essencial em áreas de alto impacto social. O Direito do Consumidor, por exemplo, ocupa uma posição central no judiciário brasileiro. De acordo com [ConJur 2023], 25% de todas as ações distribuídas nas justiças estadual e federal entre 2018 e 2022 tratavam de demandas consumeristas.

Apesar dessa relevância prática, faltam recursos de IA voltados para a área. A ausência de conjuntos de dados dificulta a avaliação da capacidade dos modelos em interpretar normas e resolver casos práticos regidos pelo Código de Defesa do Consumidor (CDC)[Brasil 1990]. Para avaliar esse conhecimento jurídico, uma das referências mais confiáveis no Brasil é o Exame de Ordem Unificado (OAB). A primeira fase do exame utiliza questões de múltipla escolha criadas justamente para testar o conhecimento legal dos candidatos.

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 573–577
October 19-22, 2026 ©2026 Association for Computational Linguistics

A fim de mitigar essa escassez de recursos em PLN, este artigo introduz o **OABench-Consumidor**, um *benchmark* composto apenas por questões de Direito do Consumidor da OAB, abrangendo as edições de 2010.1 a 2025.2. Além de detalhar a construção da base, o trabalho apresenta um experimento comparando o desempenho de um modelo fundacional com uma versão ajustada (*fine-tuned*) para atestar a sensibilidade do conjunto de dados em avaliações de domínio.

2. Trabalhos Relacionados

No Brasil, trabalhos recentes têm buscado reduzir a escassez de datasets especializados para o Direito do Consumidor. Athaydes et al. [Athaydes et al. 2024] propuseram uma metodologia para a construção de um dataset sintético de Perguntas, Respostas e Contextos a partir dos artigos do Código de Defesa do Consumidor (CDC)[Brasil 1990], Súmulas e Acórdãos extraídos do Superior Tribunal de Justiça (STJ). Em continuidade, Athaydes et al. [Athaydes et al. 2025] ampliaram essa proposta ao apresentar uma segunda versão do dataset (V2) e uma avaliação comparativa em relação à versão inicial (V1), discutindo critérios de qualidade do recurso gerado. O modelo ajustado com a V2 é retomado neste artigo na etapa experimental.

Complementando os datasets voltados à adaptação de modelos, benchmarks jurídicos também têm sido propostos para avaliar seu desempenho em tarefas específicas. Exames padronizados de admissão à advocacia, como o *Uniform Bar Examination* (UBE), têm sido usados para testar a capacidade de modelos em interpretar enunciados legais e selecionar respostas juridicamente adequadas. No Brasil, Pires et al. [Pires et al. 2026], da Maritaca AI, propuseram uma metodologia para avaliar LLMs no Exame da OAB, com foco nas questões dissertativas da segunda fase. Este artigo segue a direção de usar a OAB como fonte de avaliação, mas concentra-se nas questões objetivas da primeira fase, com recorte específico em Direito do Consumidor.

No Hugging Face [Hugging Face], também é possível encontrar repositórios com questões da primeira fase da OAB. Durante a inspeção desses repositórios públicos, observou-se ausência de recorte temático confiável para Direito do Consumidor, o que dificulta a avaliação de desempenho em áreas específicas do Direito. O OABench-Consumidor endereça essa limitação ao estruturar um benchmark voltado exclusivamente a questões objetivas de Direito do Consumidor, permitindo avaliar modelos de linguagem em um recorte jurídico mais controlado e diretamente associado ao CDC.

3. Metodologia

Esta seção descreve o processo de construção do *benchmark* OABench-Consumidor, desde a coleta até a organização final dos dados para avaliação automática de modelos de linguagem.

3.1. Coleta e Estruturação de Dados (*Web Scraping*)

O conjunto de dados foi construído a partir de questões da primeira fase do Exame de Ordem Unificado, considerando as edições de 2010.1 a 2025.2. As questões foram extraídas da plataforma *Jurisway*[1], que disponibiliza um histórico de provas da OAB organizado por edição.

A coleta foi realizada por meio de um *script* em Python, utilizando as bibliotecas `requests` e `BeautifulSoup` para acessar as páginas, interpretar o conteúdo HTML e extrair as informações necessárias. O processo seguiu três etapas principais:

1. **Mapeamento das provas:** Inicialmente, o *script* acessa a página de concursos da plataforma, identifica os *links* correspondentes ao Exame de Ordem Unificado e extrai metadados como título, edição e ano da prova.

[1] `https://www.jurisway.org.br/concursos/provas_oab.asp`

2. **Filtragem temática:** Em cada prova identificada, o *script* localiza a seção correspondente à disciplina "Direito do Consumidor" e coleta os *links* das questões associadas a essa área. Para reduzir o risco de sobrecarga no servidor, foi adotado um intervalo de 0,5 segundo entre as requisições.
3. **Extração do conteúdo:** Para cada questão selecionada, foram extraídos o enunciado e as alternativas de múltipla escolha (A, B, C e D). Como o gabarito é disponibilizado em uma página separada no *Jurisway*, o *script* também acessa o *link* de resolução, identificado como "Apenas ver Gabarito", para recuperar a alternativa correta.

3.2. Consolidação e Exportação

Após a extração, os dados foram consolidados em formato tabular utilizando a biblioteca `pandas`. O *benchmark* foi estruturado com as seguintes colunas: *Título*, *Edição*, *Data*, *Pergunta*, *Alternativas* e *Gabarito*. Por fim, o *benchmark* foi exportado para o Hugging Face[2]

4. Experimentos e Resultados

Esta seção apresenta a composição final do OABench-Consumidor e os resultados da avaliação empírica. Inicialmente, são descritos os modelos de linguagem selecionados e o *prompt* utilizado. Na sequência, detalha-se a caracterização geral do *benchmark*. Por fim, compara-se o desempenho do modelo Qwen3-8B, em sua versão base e em sua versão ajustada, na resolução das questões.

4.1. Avaliação de Modelos

Para demonstrar o uso do OABench-Consumidor como instrumento de avaliação, foi conduzido um experimento comparando duas versões do modelo Qwen3-8B:

- **Qwen3-8B (Base):** modelo fundacional avaliado sem ajuste específico para o domínio jurídico.
- **Qwen3-8B (*Fine-Tuned*):** versão do mesmo modelo ajustada para o Direito do Consumidor [Athaydes et al. 2025]. O modelo foi ajustado com um conjunto sintético de perguntas e respostas baseado em artigos do CDC, Súmulas e Acórdãos relacionados à área.

Para padronizar a avaliação, ambos os modelos receberam a mesma instrução, apresentada na Tabela 1, solicitando apenas a letra correspondente à alternativa correta.

Table 1. Estrutura do *prompt* utilizado na avaliação dos modelos.

Prompt de Avaliação
Sua tarefa é, dada a questão de múltipla escolha abaixo, assinale a letra correspondente à resposta correta, levando em consideração o Código de Defesa do Consumidor.
Pergunta: {pergunta}
{alternativas_formatadas}
Escolha a resposta correta replicando APENAS a letra correspondente à resposta correta (ex: "A").
RETORNE APENAS A ALTERNATIVA CORRETA, SEM MAIS NENHUM COMENTÁRIO.

4.2. Caracterização do *benchmark*

A coleta resultou no *benchmark* OABench-Consumidor, composto por 91 questões válidas e formatadas, abrangendo as edições de 2010.1 a 2025.2. A Tabela 2 apresenta as estatísticas gerais do conjunto. Os textos possuem em média 105 *tokens* por enunciado e 120 *tokens* nas alternativas, considerando uma palavra como um token.

[2]Link omitido para garantir o anonimato.

A distribuição dos gabaritos é apresentada na Tabela 3. As alternativas corretas aparecem de forma relativamente equilibrada, variando de 20,9% para a letra B a 28,6% para a letra C. Esse resultado reduz a possibilidade de estratégias baseadas apenas na frequência das alternativas e reforça a adequação do conjunto para avaliação objetiva de modelos.

Table 2. Estatísticas gerais do *benchmark*.

Métrica	Valor
Total de Questões	91
Média de tokens (Enunciado)	104,68
Média de tokens (Alternativas)	120,43

Table 3. Distribuição dos gabaritos no *benchmark*.

Alternativa	Frequência	Percentual
A	21	23,1%
B	19	20,9%
C	26	28,6%
D	25	27,5%

4.3. Comparação dos Modelos

Após a caracterização do conjunto, o *benchmark* OABench-Consumidor foi utilizado para comparar o desempenho dos modelos selecionados. A Tabela 4 apresenta o número de acertos e a acurácia obtida por cada modelo nas 91 questões do *benchmark*.

Table 4. Desempenho dos modelos no OABench-Consumidor.

Modelo	Acertos	Acurácia
Qwen3-8B (Base)	47 de 91	51,65%
Qwen3-8B (*Fine-Tuned*)	59 de 91	64,84%

O modelo base acertou 47 questões (51,65%). O modelo submetido ao ajuste fino com perguntas e respostas sobre os artigos do CDC e jurisprudências alcançou 59 acertos (64,84%). Esse aumento demonstra que a base construída é sensível à adaptação de domínio e pode ser útil como uma ferramenta precisa de avaliação.

Além da diferença na acurácia, também foram observadas diferenças qualitativas no comportamento dos modelos. Durante a execução, o modelo base apresentou tendência a descumprir as restrições do *prompt*, gerando respostas longas ou caracteres desconexos em vez de retornar apenas a alternativa escolhida. Foram realizados testes limitando a quantidade de *tokens* de saída para tentar forçar respostas mais curtas, mas, mesmo assim, o modelo base muitas vezes manteve a tendência de descumprir as restrições impostas. Em contrapartida, o modelo ajustado (*fine-tuned*) demonstrou forte alinhamento à instrução, respondendo de forma objetiva e retornando estritamente apenas a letra da alternativa solicitada.

5. Conclusão

A avaliação de modelos de linguagem no Direito brasileiro exige recursos padronizados e atualizados. Para suprir essa falta, este artigo apresentou o OABench-Consumidor, um *benchmark* focado em Direito do Consumidor, contendo 91 questões de edições históricas da OAB.

A aplicação prática do conjunto validou sua utilidade. A avaliação demonstrou que o modelo Qwen3-8B ajustado para o contexto do CDC superou a versão base, elevando a acurácia de 51,65% para 64,84%. Esses resultados evidenciam que o *benchmark* consegue capturar de forma objetiva a evolução do conhecimento jurídico dos modelos avaliados.

Agradecimentos

Os autores agradecem o financiamento parcial através da bolsa UFBA PIBIC 2025/2026, do apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001, FAPESB CCE 0022/2023, FAPESB TIC 002/2015. Este trabalho também conta com o apoio do Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação (INCT-TILD-IAR) (processo 408490/2024-1). Gostaríamos de agradecer ao Escavador pela parceria neste trabalho.

Referências

Athaydes, A., Bulcao, L., Sacramento, C., Mane, B., Claro, D. B., Souza, M., and Pita, R. (2024). Brazilian consumer protection code: a methodology for a dataset to question-answer (QA) models. In Claro, D. B. and Pagano, A., editors, *Proceedings of the 15th Brazilian Symposium in Information and Human Language Technology*, pages 522–529, Belém do Pará, Brazil. Association for Computational Linguistics.

Athaydes, A., Mota, L., Neto, F. M., Silva, S., Mane, B., Claro, D., Souza, M., and Lisboa, A. (2025). Towards a corpus methodology for llms in the legal domain. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 272–282, Porto Alegre, RS, Brasil. SBC.

Brasil (1990). Lei nº 8.078, de 11 de setembro de 1990. Dispõe sobre a proteção do consumidor e dá outras providências. Diário Oficial da União, Brasília, DF.

ConJur (2023). De cada 4 ações nas justiças estadual e federal, uma é de consumo. `https://www.conjur.com.br/2023-out-11/cada-acoes-justicas-estadual-federal-consumo/`. Acesso em: 21 jun. 2025.

Hugging Face. Hugging face: The ai community building the future. `https://huggingface.co`. Acesso em: março 2026.

Pires, R., Malaquias Junior, R., and Nogueira, R. (2026). Automatic legal writing evaluation of llms. In *Proceedings of the Twentieth International Conference on Artificial Intelligence and Law*, ICAIL '25, pages 420–424, New York, NY, USA. Association for Computing Machinery.

Análise de comportamento do modelo NotebookLM em transcrição automática

Brenda R. F. de Oliveira[1], **Oto Araújo Vale**[1]

[1]Departamento de Letras – Universidade Federal de São Carlos (UFSCar)
13565-905 – São Carlos – SP – Brasil

{brendafernandes@estudante.ufscar.br, otovale@ufscar.br}

Abstract. *This research analyzes the behavior of the NotebookLM model in automatic speech transcription, using a corpus of 55 economic interviews from the Roda Viva TV program. The study is based on the hypothesis that the model's textual registry choices reveal both its capabilities and limitations regarding spontaneous speech. By identifying and organizing behavioral patterns, a typology representing the model's behavioral particularities was constructed to form an annotation guideline. Results obtained so far demonstrate that the model alternates between sensitivity to speech disfluencies and limitations in textual organization.*

Resumo. *Esta pesquisa analisa o comportamento do modelo NotebookLM na transcrição automática da fala, tomando como corpus a transcrição de 55 entrevistas de economia do programa Roda Viva. O estudo parte da hipótese de que as escolhas de registro textual do modelo revelam tanto suas capacidades quanto suas limitações diante da fala espontânea. Com a identificação e organização de padrões de comportamento, construiu-se uma tipologia que representa as particularidades de comportamento do modelo para compor uma diretriz de anotação. Os resultados obtidos até o momento demonstram que o modelo alterna entre sensibilidade às disfluências da fala e limitações na organização textual.*

1. Quadro teórico e objetivo de pesquisa

A presente pesquisa tem como objetivo analisar sistematicamente o comportamento do modelo NotebookLM[1] na tarefa de transcrição automática da fala, tomando como base um conjunto de entrevistas de economia do programa Roda Viva da TV Cultura. Nesse sentido, o interesse central reside em compreender as tendências comportamentais e limitações do sistema computacional, ao identificar padrões recorrentes em suas produções textuais, avaliar escolhas feitas pelo modelo e a forma como ele lida com a natureza complexa da fala. Tomou-se como premissa de que, além de ser caracterizada por uma natureza dinâmica e fragmentária [Clark and Fox Tree 2002, Galembeck 1999], a fala espontânea apresenta caráter quase simultâneo na relação entre seu planejamento e a sua realização [Duarte et al. 2014]. Nesse contexto, a fala não é compreendida como produto acabado, mas como processo cuja materialidade permite que existam marcas de planejamento em sua superfície e estrutura, conhecidas como disfluências [Madureira 2026].

[1]notebooklm.google.com

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 578–582
October 19-22, 2026 ©2026 Association for Computational Linguistics

A relevância crescente das tecnologias baseadas em fala, como assistentes de voz, sistemas de reconhecimento automático de fala (ASR) e plataformas de diálogo, demonstra que a transcrição automática ocupa papel fundamental. Isso se deve a qualidade das interações entre humano e máquina que depende, em grande medida, da precisão com que os sinais sonoros são convertidos em linguagem escrita. No entanto, a fala espontânea representa um desafio particular para Grandes Modelos de Linguagem Multimodais (MLLM), pois é mais complexa de processar e transcrever com precisão, [Casanova et al. 2026], além de carecer de estrutura clara, ser fragmentada e apresentar disfluências. Com isso, parte-se do pressuposto de que o comportamento do modelo, materializado nas escolhas de registro textual, pode revelar suas limitações técnicas, bem como sua capacidade de lidar com a complexidade constitutiva da fala em interação.

2. Metodologia

A metodologia utiliza abordagem qualitativa e descritiva, com o apoio da quantificação das ocorrências observadas. Nesse sentido, partiu-se de um *corpus* composto pela transcrição de 55 entrevistas de economia do programa Roda Viva que foram realizadas entre 2016 e 2026 e estão disponíveis no canal oficial do programa no YouTube[2]. Tratam-se de produções de longa duração em língua portuguesa brasileira e de livre acesso ao público. Pesquisas recentes utilizam *corpora* ideais para analisar a fala espontânea autêntica, como o NURC-SP [Lima et al. 2024] e o CORAA ASR [Gris et al. 2022], e avaliar modelos como Wav2Vec2 e Whisper por meio de métricas padrão como a *Word Error Rate* (WER). Diferentemente desses *corpora*, reconhece-se que as entrevistas, relacionadas ao *corpus* do presente trabalho, não representam integralmente a fala espontânea, pois o discurso tende a apresentar formalidade e monitoramento acentuados.

Entretanto, tais contextos de interação preservam características da fala espontânea como hesitações, reformulações, repetições, assim como a simultaneidade entre planejamento e produção. Isso aproxima tais interações de circunstâncias de fala em uso e as diferem de situações de oralização da escrita[3], tornando-as adequadas para a análise deste estudo. Além disso, a métrica WER avalia o desempenho lexical de sistemas de ASR com a normalização de transcrições e cálculo. O presente trabalho, no entanto, propõe uma investigação para além da avaliação mecânica, ao preservar as transcrições e incluir abordagens qualitativas, considerando aspectos[4] como a segmentação, a capitalização, a pontuação, além do registro de fenômenos da fala espontânea.

O arquivo de áudio de cada entrevista foi importado para a plataforma NotebookLM que realizou automaticamente a conversão dos sinais sonoros em texto escrito sem oferecer a opção para incluir prompt. Com isso, não foi fornecido nenhuma instrução adicional, permitindo preservar o funcionamento original do modelo. As transcrições resultantes foram organizadas em um editor de texto para análise. Na etapa de identificação manual de padrões de comportamento do modelo, as ocorrências foram sinalizadas no

[2] `www.youtube.com/rodaviva`

[3] A leitura em voz alta e a realização de discursos roteirizados são alguns exemplos de oralização da escrita.

[4] Embora não sejam contemplados diretamente pela métrica WER, acredita-se que tais fatores também influenciam a qualidade e a precisão da transcrição tanto no âmbito lexical quanto na interpretação da informação.

próprio corpo do texto e classificadas segundo sua natureza. Nesse processo, foi realizada a comparação manual das transcrições com o auxílio do vídeo das respectivas entrevistas, permitindo verificar a equivalência entre o conteúdo verbal e o texto escrito. Em casos de incerteza, os trechos correspondentes foram reproduzidos novamente, a fim de confirmar a classificação atribuída. As ocorrências anotadas foram registradas em planilhas por meio de ferramentas automatizadas, para a contagem das frequências e a verificação sistemática dos padrões identificados. Essa organização permitiu desenvolver uma tipologia que representa as particularidades de comportamento do modelo para compor uma diretriz que orienta a anotação das demais transcrições presentes no *corpus*.

3. Resultados e discussão

Até o momento, foram identificadas dez categorias, distribuídas em três grandes grupos, sendo a primeira voltada para o registro de marcas de disfluência e planejamento da fala; o segundo para os desvios de normatização e escrita; e o terceiro voltado para inadequações estruturais e de conteúdo. No primeiro grupo, foram observadas categorias como (i) Registro de repetição, (ii) Registro de hesitação e (iii) Registro de reformulação da fala. No segundo, incluem-se (i) Desvio ortográfico, cujas subcategorias são (i.a) Palavra e (i.b) Número; (ii) Desvio de pontuação[5] e (iii) Desvio de capitalização. Já o terceiro grupo abrange (i) Inadequação de segmentação[6], (ii) Omissão[7], (iii) Inserção e Substituição lexical, sendo esta última subdividida em (i.a) Por homofonia e (i.b) Por alucinação[8].

Tendo isso em vista, entende-se que as categorias presentes tanto no primeiro quanto no segundo grupo revelam as limitações técnicas e as escolhas do modelo na tarefa de transcrever a fala. A análise inicial indica que ele demonstrou dificuldade em aplicar as normas da escrita a um material da fala que possui especificidades próprias. Já as categorias presentes no primeiro grupo refletem a sensibilidade do modelo às disfluências naturais da fala. Isso pode ser observado no registro de hesitação em que o modelo reconhece o fenômeno e representa, em diversos casos, como "eh" em vez de "é". A frequência que as categorias apresentam são variadas e podem ser observadas com os dados da transcrição da entrevista de Ricardo Alban[9] como exemplo presente na tabela 1.

Nessa entrevista, os dados mostram que os desvios ortográficos e registros de repetição aparecem com maior incidência, seguidos por registros de hesitação e desvios de capitalização. Isso sugere que o modelo alterna entre sensibilidade às disfluências da fala e falhas na organização textual. Nessa análise inicial, o modelo também demonstrou falhar em capturar certos trechos da fala humana como, por exemplo, a fala "O ministro Fernando Haddad [não dá os detalhes da medida], mas já adiantou que boa parte delas pre-

[5]Nessa categoria, considera-se as ocorrências com uso de sinais gráficos em desacordo com as normas gramaticais. Esse desvio pode ocorrer pelo registro de determinada pontuação em situação que não permite sua inclusão ou pela ausência de pontuação.

[6]Nessa categoria, considera-se as ocorrências em que palavra ou trecho é escrito com separação em desacordo com a norma-padrão, resultando em uma ocorrência de hipossegmentação ou hipersegmentação.

[7]A omissão ocorre quando parte do discurso gravado não é registrada na transcrição. Ocorrências caracterizadas pela ausência de letras dentro de uma palavra não fazem parte dessa categoria, mas da "Desvio ortográfico".

[8]Considera-se como alucinação, quando o modelo substitui palavras, presentes no áudio original, por outras que não foram ditas.

[9]Presidente da Confederação Nacional da Indústria (CNI), entrevista realizada em 04/08/2025 - `https://www.youtube.com/watch?v=EjOvy76WA2M`

Tabela 1. Frequência de ocorrência por categoria

Natureza da categoria	Categorias	Frequência
Marcas de disfluência e planejamento da fala	Registro de repetição	121
	Registro de hesitação	93
	Registro de reformulação da fala	51
Desvios de normatização e escrita	Desvio ortográfico	169
	Palavra	167
	Número	2
	Desvio de pontuação	41
	Desvio de capitalização	82
Inadequações estruturais e de conteúdo	Inadequação de segmentação	27
	Omissão	63
	Inserção	14
	Substituição lexical	41
	Por homofonia	36
	Por alucinação	5

cisaria passar pelo Congresso Nacional de um jeito ou de outro, né?", em que o fragmento delimitado por colchetes estava presente na fala gravada, mas foi omitido na transcrição pelo modelo. Nesse sentido, o NotebookLM não apenas reconheceu certos aspectos da oralidade, mas também impôs uma reorganização linguística sobre a fala que, nesse caso, comprometeu a interpretação do que foi dito.

4. Conclusão

Este estudo une Linguística e Processamento de Língua Natural. Assim, espera-se não apenas descrever padrões de comportamento do sistema, mas também contribuir para avaliações mais refinadas de tecnologias de transcrição automática no português brasileiro. As limitações manifestadas pelo NotebookLM demonstram que ele, enquanto um MLLM, apresentou dificuldade em aplicar automaticamente as regras formais da escrita na transcrição da fala humana. Por outro lado, diferente de sistemas de conversão mecânica de sinais sonoros em texto, o modelo NotebookLM revelou sensibilidade ao planejamento da fala, com a inferência contextual necessária para reconhecer e distinguir tanto hesitações como "eh", de formas verbais como "é", quanto às demais disfluências, reforçando habilidades de reorganização linguística.

Considerando que esta pesquisa ainda está em desenvolvimento, cabe ressaltar que os resultados apresentados possuem caráter preliminar, pois estão voltados apenas para a etapa inicial da investigação e referem-se exclusivamente ao NotebookLM. Pretende-se realizar posteriormente uma comparação com os resultados obtidos pelo modelo Whisper para verificar se os padrões observados são específicos do NotebookLM ou compartilhados por outros sistemas. Nesse sentido, as discussões realizadas não buscam propor conclusões definitivas sobre o conjunto de entrevistas, mas ilustrar os padrões observados até o momento. Portanto, eventuais generalizações poderão ser realizadas somente após a conclusão da análise integral do *corpus* e da etapa comparativa.

Referências

Casanova, E., Santos, V. G., Svartman, F. R. F., Leite, M. Q., Candido Junior, A., Marcacini, R. M., Rezende, S. O., and Aluísio, S. M. (2026). Recursos para o processamento de fala. In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, volume 1, book chapter 3. BPLN, 4 edition.

Clark, H. H. and Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111.

Duarte, T. P. P. J., Storto, L. J., and Durante, D. (2014). Características da língua falada e da língua escrita: análise de entrevistas com vik muniz.

Galembeck, P. d. T. (1999). Metodologia de pesquisa em português falado. In Rodrigues, A. C. d. S., Alves, I. M., and Goldstein, N. S., editors, *I Seminário de Filologia e Língua Portuguesa*, pages 109–119. Humanitas Publicações FFLCH/USP, São Paulo.

Gris, L., Junior, A. C., Santos, V., Dias, B., Leite, M., Svartman, F., and Aluísio, S. (2022). Bringing nurc/sp to digital life: the role of open-source automatic speech recognition models. In *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional*, pages 330–341, Porto Alegre, RS, Brasil. SBC.

Lima, R., Leal, S., Junior, A. C., and Aluísio, S. (2024). A large dataset of spontaneous speech with the accent spoken in são paulo for automatic speech recognition evaluation. In *Anais da XXXIV Brazilian Conference on Intelligent Systems*, pages 33–47, Porto Alegre, RS, Brasil. SBC.

Madureira, B. (2026). Diálogo e interatividade. In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, volume 1, book chapter 15. BPLN, 4 edition.

Desafios Tipológicos na Construção de Treebanks: O Mapeamento Morfossintático da Língua Kubeo no Padrão UD

Pedro Paulo Freire Oliveira[1], Thiago Blanch Pires[2]

[1 2]Instituto de Letras – Universidade de Brasília, Campus Universitário Darcy Ribeiro, Gleba A, Asa Norte – CEP 70910-900 – Brasília – DF – Brasil

pedro.freire@aluno.unb.br, thiagocomaga@gmail.com

Abstract. *The inclusion of Amazonian indigenous languages in the field of Natural Language Processing (NLP) is hindered by the lack of treebanks and structured resources. This paper describes the initial phase of morphosyntactic modeling of the Kubeo language (Eastern Tukanoan family) under the Universal Dependencies (UD) guidelines. Kubeo exhibits high typological complexity, characterized by agglutinative processes, pervasive nominal classifiers, and pragmatic markers that challenge the static grammatical categories of the UD standard. The scope of this stage focuses on the extraction and tagging of native sentences, with a methodological emphasis on measuring the structural distance between the explicit nominal subject and the core verb. Based on data from the Pamiene Mama Buei-Tukubo primer, the study proposes token mapping solutions, utilizing the morphological features layer FEATS to preserve the language's morphosyntactic richness without excessively fragmenting the dependency trees. This pioneering corpus aims to establish a reproducible annotation protocol for future Kubeo computational tools.*

Resumo. *A inserção de línguas indígenas amazônicas no campo do Processamento de Linguagem Natural (PLN) esbarra na ausência de treebanks e recursos estruturados. Este artigo descreve a fase inicial de modelagem morfossintática da língua Kubeo (família Tukano-Oriental) sob as diretrizes das Dependências Universais (UD). O Kubeo apresenta alta complexidade tipológica, caracterizada por processos aglutinantes, classificadores nominais pervasivos e marcadores pragmáticos que desafiam as categorias gramaticais estáticas do padrão UD. O escopo desta etapa concentra-se na extração e etiquetagem de sentenças nativas, com foco metodológico na mensuração da distância estrutural entre o sujeito nominal explícito e o verbo núcleo. A partir de dados da cartilha Pamiene Mama Buei-Tukubo, o estudo propõe soluções de mapeamento de tokens, utilizando a camada de traços morfológicos FEATS para preservar a riqueza morfossintática da língua sem fragmentar excessivamente as árvores de dependência. A criação deste corpus visa estabelecer um protocolo de anotação reprodutível para futuras ferramentas focadas no Kubeo.*

1. Introdução

A construção de recursos computacionais para línguas de baixo recurso constitui um dos principais desafios da Linguística Computacional, especialmente no caso das línguas indígenas brasileiras, para as quais ainda há uma carência crítica de corpora morfossintaticamente anotados e *treebanks*, limitando o desenvolvimento de ferramentas de Processamento de Linguagem Natural (ALENCAR, 2021; ALEXANDRE et al., 2021; CASELI; NUNES, 2022)

Entre essas línguas, o Kubeo, pertencente à família Tukano-Oriental e falado na região do alto rio Uaupés, apresenta características tipológicas que representam desafios relevantes para a anotação computacional, como morfologia aglutinante, evidencialidade

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 583–587
October 19-22, 2026 ©2026 Association for Computational Linguistics

gramaticalizada, classificadores nominais e relativa flexibilidade sintática. Embora a língua conte com uma descrição linguística consolidada, sobretudo nos trabalhos de Chacon (2012), ainda não dispõe de um *treebank* compatível com o padrão *Universal Dependencies* (UD), o que restringe sua integração ao ecossistema internacional de recursos linguísticos.

Este trabalho integra um projeto de pesquisa voltado ao desenvolvimento de recursos computacionais para línguas indígenas brasileiras e utiliza como corpus inicial sentenças extraídas da cartilha *Pamiene Mama Buei-Tukubo* (2023), produzida no âmbito das ações de documentação e ensino da língua Kubeo conduzidas por Thiago Chacon e colaboradores. Nessa perspectiva, o objetivo desta etapa é estabelecer uma metodologia consistente de anotação morfossintática, capaz de subsidiar a expansão do corpus e o desenvolvimento de ferramentas de PLN para a língua.

O processo de anotação adota como referência o modelo *Universal Dependencies* (Nivre et al., 2016; de Marneffe et al., 2021) atualmente um dos principais padrões internacionais para representação sintática de corpora anotados. Entretanto, sua aplicação a línguas tipologicamente distantes das línguas indo-europeias levanta desafios metodológicos conhecidos na literatura brasileira como Problema de Anchieta (SANDALO; GALVES, 2017), relacionados às dificuldades de representar sistemas morfossintáticos tipologicamente distintos por meio de categorias gramaticais universais. No caso do Kubeo, esses desafios concentram-se principalmente na representação de morfemas multifuncionais e na necessidade da utilização extensiva da camada de traços morfológicos FEATS, seguindo as recomendações do próprio projeto UD.

Diante desse cenário, este artigo apresenta a etapa inicial da construção de um *treebank* para a língua Kubeo compatível com UD São descritos o protocolo inicial de anotação desenvolvido para um conjunto piloto de sentenças, as principais decisões metodológicas adotadas durante o processo de anotação e uma análise preliminar da distância estrutural entre sujeito nominal explícito e verbo núcleo. Ao sistematizar esses procedimentos, o trabalho estabelece uma base metodológica para a ampliação do corpus e para o desenvolvimento de recursos computacionais destinados ao PLN da língua Kubeo.

2. Fundamentação Teórica

Universal Dependencies (UD) é um esquema de anotação morfossintática concebido para representar diferentes línguas por meio de um conjunto padronizado de categorias gramaticais, relações de dependência e traços morfológicos, favorecendo a interoperabilidade entre corpora e o desenvolvimento de ferramentas multilíngues de Processamento de Linguagem Natural (de Marneffe et al., 2021). O modelo organiza as anotações no formato CoNLL-U, estruturado em dez colunas que registram informações como forma, lema, classe gramatical, traços morfológicos e relações sintáticas, permitindo a construção de *treebanks* compatíveis entre diferentes projetos linguísticos.

Nos últimos anos, o UD tem sido progressivamente aplicado a línguas de baixo recurso, incluindo algumas línguas indígenas brasileiras, demonstrando o potencial do modelo para apoiar tanto estudos tipológicos quanto aplicações computacionais. Entretanto, o número de *treebanks* disponíveis ainda é reduzido quando comparado ao de línguas amplamente documentadas, tornando a ampliação desses recursos uma demanda relevante para a Linguística Computacional (Alexandre et al., 2021; Alencar, 2021).

A aplicação de um modelo universal a línguas tipologicamente distintas suscita, contudo, desafios de representação. Sandalo e Galves (2017) denominam esse cenário de Problema de Anchieta, referindo-se às dificuldades de descrever línguas indígenas americanas por meio de categorias gramaticais originalmente formuladas para línguas indo-europeias. No caso do Kubeo, essas dificuldades manifestam-se especialmente na

segmentação de morfemas multifuncionais e na representação de informações morfológicas e discursivas que nem sempre encontram correspondência direta nas categorias universais do UD.

Diante dessas limitações, este trabalho adota a camada de traços morfológicos FEATS prevista na própria especificação do UD para representar informações gramaticais que não podem ser expressas apenas pelas categorias lexicais e pelas relações de dependência. Essa escolha metodológica não altera a arquitetura do modelo, mas explora um de seus mecanismos de extensão para preservar distinções morfossintáticas relevantes da língua Kubeo durante o processo de anotação.

3. Metodologia

O corpus inicial é composto por um conjunto de 20 sentenças extraídas da cartilha *Pamiene Mama Buei-Tukubo* (2023). O material foi selecionado por apresentar construções representativas da língua e por possuir revisão linguística prévia, constituindo uma base adequada para o desenvolvimento do protocolo inicial de anotação. As sentenças foram selecionadas de forma gradual, seguindo um critério de complexidade sintática crescente. Inicialmente, priorizaram-se construções mono-oracionais, expandindo-se posteriormente para estruturas mais complexas à medida que o protocolo de anotação era desenvolvido e as categorias morfossintáticas eram refinadas para a representação da língua Kubeo. Esse processo foi conduzido de forma iterativa, com validações globais e revisões sucessivas das anotações, a fim de assegurar a consistência dos critérios adotados ao longo da construção do corpus.

A anotação foi realizada manualmente no formato *CoNLL-U*, seguindo as diretrizes do UD. Para cada sentença, foram realizadas a segmentação e a lematização, bem como a anotação das categorias morfossintáticas (UPOS), dos traços morfológicos FEATS, das relações sintáticas e da árvore de dependências. Nos casos em que a morfologia do Kubeo não encontrava correspondência direta nas categorias universais, adotaram-se os mecanismos de representação previstos pelo próprio UD, preservando informações relevantes por meio dos traços morfológicos. O protocolo foi refinado iterativamente durante a anotação, de modo a garantir coerência na representação de fenômenos morfossintáticos recorrentes do Kubeo.

Como estudo de caso para avaliação do corpus piloto, analisou-se a distância linear entre sujeitos nominais explícitos e seus respectivos verbos núcleo. Essa métrica foi selecionada por refletir diretamente a estrutura sintática anotada e por permitir verificar a uniformidade das relações de dependência produzidas durante o processo de anotação, além de fornecer um primeiro diagnóstico quantitativo da organização sintática do corpus.

4. Resultados

Como resultado desta etapa, foi produzido um corpus piloto anotado segundo o padrão UD, contemplando segmentação, lematização, categorias morfossintáticas, traços morfológicos e relações de dependência. O processo de anotação permitiu consolidar um protocolo consistente para a representação de fenômenos recorrentes da língua Kubeo, estabelecendo uma base metodológica para a expansão futura do *treebank* e para o desenvolvimento de recursos computacionais voltados ao Processamento de Linguagem Natural.

Como estudo de caso estrutural, analisou-se a distância linear entre sujeitos nominais explícitos e seus respectivos verbos núcleo. O processamento das 20 sentenças do corpus piloto demonstrou a presença do sujeito nominal (nsubj) em 90% da amostra, com forte predominância da ordem Sujeito-Verbo. A distância linear variou de 1 a 4 tokens, com média de 2,5. O distanciamento máximo atingido é advindo da presença de constituintes intervenientes, como pode ser observado na Tabela 1, além de registrar

flexibilidade sintática nos casos em que o verbo precede o sujeito. Esses resultados evidenciam que o protocolo adotado foi capaz de representar adequadamente tanto dependências locais adjacentes quanto dependências de maior extensão.

Tabela 1. Anotação morfossintática no formato CoNLL-U para a sentença 1: "Yawi **oko hãrãwʉi ãino wore kuibi**" (A onça procura comida no dia de chuva).

ID	Forma	Morfemas	UPOS	Feats	DepRel	Head
1	Yawi	yawi (onça)	NOUN	Animacy=AnimlGender=MasclNumber=Sing	nsubj	5
2	oko	oko (água/chuva)	NOUN	Animacy=InanlNumber=Sing	nmod	3
3	hãrãwʉi	hãrãwʉ (dia) -i (LOC)	NOUN	Animacy=InanlCase=Loc	obl	5
4	ãino	ãino (comida)	NOUN	Animacy=InanlCase=AcclNumber=Sing	obj	5
5	wore	wo (procurar) -re	VERB	VerbForm=Vnoun	root	0
6	kuibi	kui (estar/AUX) -bi (3M)	AUX	Gender=MasclNumber=SinglPerson=3	aux	5

Durante a anotação, verificou-se que determinados fenômenos morfossintáticos do Kubeo demandaram decisões interpretativas que extrapolam uma aplicação estritamente mecânica das diretrizes do UD. A utilização intensiva da camada FEATS permitiu preservar distinções linguísticas relevantes sem comprometer a compatibilidade do corpus com o padrão internacional, evidenciando que, embora o UD forneça uma estrutura robusta para a anotação, a aplicação a línguas tipologicamente distintas daquelas que fundamentaram seu desenvolvimento requer adaptações metodológicas bem documentadas.

Em conjunto, os resultados demonstram a viabilidade da construção incremental de um treebank para a língua Kubeo compatível com UD. Além da produção de um corpus piloto anotado, o trabalho estabeleceu um protocolo reprodutível que poderá orientar a expansão do corpus e subsidiar o desenvolvimento de ferramentas computacionais para a documentação e o processamento automático da língua.

5. Conclusões e Trabalhos Futuros

Este trabalho apresentou a primeira etapa da construção de um treebank para a língua Kubeo compatível com o padrão UD, estabelecendo um protocolo de anotação morfossintática desenvolvido a partir de um corpus piloto. Os resultados demonstraram a viabilidade da metodologia proposta e evidenciaram que a utilização dos mecanismos previstos pelo próprio UD, especialmente os traços morfológicos FEATS, permite representar fenômenos relevantes da língua Kubeo preservando a compatibilidade com o padrão internacional.

Como continuidade da pesquisa, pretende-se ampliar o corpus anotado com foco em materiais didáticos e contextos de ensino, incorporando as demais sentenças da cartilha *Pamiene Mama Buei-Tukubo* (2023) e outras fontes desse domínio, a fim de consolidar o *treebank* da língua Kubeo e disponibilizá-lo como recurso aberto para a comunidade científica. Esse recurso servirá de base para o desenvolvimento de ferramentas de Processamento de Linguagem Natural, como etiquetadores morfossintáticos e analisadores sintáticos, que, após validação supervisionada, poderão apoiar a expansão semiautomática do corpus, tornando-o mais robusto e abrangente. Espera-se que esses produtos contribuam para a documentação digital, a descrição linguística e a preservação da língua Kubeo, ampliando sua representação no ecossistema de recursos computacionais.

Este estudo piloto, de natureza estritamente documental baseada em cartilhas, apresenta limitações de generalização para fenômenos complexos, que serão tratados futuramente, e não envolve intervenção com seres humanos. Declara-se o uso de IA Generativa exclusivamente para revisão gramatical e aprimoramento estilístico do texto final, conforme diretrizes da SBC.

Referências

ALENCAR, L. F. de. Uma gramática computacional de um fragmento do Nheengatu. Revista Estudos da Linguagem, Belo Horizonte, v. 29, n. 3, p. 1717–1777, 2021.

ALEXANDRE, D. M.; GURGEL, J. L.; ALENCAR, L. F. de. Nheentiquetador: um etiquetador morfossintático para o sintagma nominal do Nheengatu. In: Anais do XIII Encontro de Linguística de Corpus. 2021.

CASELI, H. de M.; NUNES, M. das G. V. (eds.). Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português. São Paulo: SBC, 2022.

CHACON, T. C. The phonology and morphology of Kubeo: the documentation, theory, and description of an Amazonian language. Tese (Doutorado em Linguística) - University of Hawai‘i at Manoa, Honolulu, 2012.

CHACON, T. C. (Org.). Pamiene Mama Buei-Tukubo: Primeiro livro para o ensino da língua Kubeo. Escola Estadual Indígena Huredanaçu Nossa Senhora da Imaculada Conceição; SEDUC-AM, 2023.

DE MARNEFFE, M.-C.; et al. Universal Dependencies. Computational Linguistics, v. 47, n. 2, p. 255–308, 2021.

NIVRE, J.; et al. Universal Dependencies v1: A multilingual treebank collection. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16). Portorož, Slovenia: European Language Resources Association, 2016.

SANDALO, M. F. S.; GALVES, C. M. C. Anotando sintaticamente uma língua originária do Brasil: o problema de Anchieta. Cadernos de Estudos Linguísticos, Campinas, v. 59, n. 3, p. 631–648, 2017.

Unsupervised Subword Segmentation for POS Tagging in Low-Resource Agglutinative Languages

Jonas Oliveira Pereira[1], **Lilian Teixeira de Sousa**[1], **Marlo Souza**[1]

[1]Federal University of Bahia (UFBA)
Campus Ondina, 40.170-110 Salvador, Brazil

{jonas.pereira, lilian.sousa, msouza1}@ufba.br

Abstract. *While Part-of-speech tagging is considered a well-understood task in the literature, most work has focused on indo-european languages with fusional morphology. For low-resource agglutinative languages, such as brazillian indigenous languages, the task is commonly constrained by small annotated corpora, high lexical sparsity, and morphological patterns that are poorly represented by word-level models. This paper investigates the impact of unsupervised subword segmentation techniques on POS tagging for brazillian indigenous languages. We compare a word-level baseline with Byte Pair Encoding, Morfessor, and FlatCat on Bororo, Nheengatu, and Tupinamba corpora, including a controlled experiment on the size of the training corpus. Our findings suggest that unsupervised segmentation can reduce sparsity in low-resource POS tagging, although its benefit depends on the language, corpus size, and segmentation method.*

1. Introduction

Agglutinative languages pose a difficult challenge for natural language processing due to its morphological richness [Tsarfaty et al. 2013, Özçift et al. 2021]. More yet, for those in low-resource settings, such as for Brazilian Indigenous languages, annotated corpora are often too small to support reliable supervised training [Mager et al. 2018]. This is especially problematic for part-of-speech (POS) tagging since a single word may contain several morphologically relevant affixes, carrying grammatical information that helps determine the word's tag. As a result, the same stem can appear in many different surface forms, increasing lexical sparsity and making rare or unseen forms more common during training and evaluation [Mager et al. 2018, Mager et al. 2022].

For Latin American Indigenous and other under-documented languages, this problem is intensified by practical constraints: digital corpora are small, linguistic expertise is scarce, and annotation resources are expensive to produce [Tonja et al. 2024, Mager et al. 2018]. A word-level POS tagger trained under these conditions may fail to generalize across related surface forms, even when those forms share recurring stems or affixes [Eskander et al. 2022, Libovický and Helcl 2024]. Unsupervised subword segmentation offers a possible way to expose these recurring units without requiring manually segmented data [Mager et al. 2022, Virpioja et al. 2013].

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 588–593
October 19-22, 2026 ©2026 Association for Computational Linguistics

This paper investigates whether unsupervised subword segmentation improves POS tagging for low-resource agglutinative languages. The remainder of the paper is organized as follows. Section 2 presents related work. Section 3 describes the segmentation methods, corpora, and experimental setup. Section 4 reports and interprets the results, and Section 5 concludes the paper.

2. Related Work

Prior work has studied subword and morphological segmentation as a way to reduce sparsity in low-resource morphologically complex languages. [Mager et al. 2022] compare BPE with supervised and unsupervised morphological segmentation for machine translation in four polysynthetic languages, finding that unsupervised morphological segmentation often outperforms BPE for the task. [Saleva and Lignos 2021] evaluate morphology-aware segmentation for low-resource neural machine translation, but find no consistent advantage over BPE across language pairs. More recently, [Libovický and Helcl 2024] propose lexically grounded segmentation methods and report clearer gains for POS tagging than for machine translation. These results suggest that segmentation effects are task-dependent and especially relevant for morphologically sensitive tasks.

For POS tagging, [Eskander et al. 2022] study cross-lingual POS tagging for low-resource morphologically rich languages by projecting tags through stems and morphemes instead of full word forms. Their results show that replacing sparse surface words with stem- or morpheme- based representations improves projection-based tagging. [Keren et al. 2022] compare character-level and subword-level pretrained models on morphologically rich languages, including POS tagging and morphological disambiguation, and find that subword models are not uniformly better across tasks.

Subword segmentation methods differ in the assumptions they make about word structure. BPE learns frequent symbol merges and is widely used to handle rare words in neural NLP [Sennrich et al. 2016]. Morfessor models unsupervised morphological segmentation from raw word forms [Virpioja et al. 2013], while FlatCat extends the Morfessor family with an HMM-based model that assigns morph categories, targeting unsupervised and semi-supervised morphology learning [Grönroos et al. 2014]. Although these methods are often evaluated in translation or segmentation tasks, their effect on POS tagging in limited resource languages, particularly Latin American Indigenous-languages corpora remains less explored.

3. Methodology

We compare four input representations for POS tagging: a word-level baseline and three subword segmented variants produced by BPE, Morfessor, and FlatCat. This design tests whether subword information helps the tagger handle sparse word forms while still predicting POS tags at the word level.

3.1. Word-level baseline

We employ a BiLSTM-CRF [Huang et al. 2015] tagger, a competitive approach for sequence tagging [Souza et al. 2019] without requiring large pretrained models. As a baseline we train and evaluate the model on the original word-level tokens, without subword segmentation. Thus, all gains reported in the results are computed against this baseline under the same dataset split, evaluation unit, and model configuration.

3.2. Corpora and Datasets

The experiments use three agglutinative languages that have publicly available datasets based on Universal Dependencies: Bororo [Gerardi et al. 2023], Nheengatu [de Alencar 2024], and Tupinamba [Gerardi 2023]. Since among the languages studied, only the Bororo corpus is sufficiently large to allow further segmentation, we conduct a study on the impact of the training corpus size in the performance, by selecting segments from the corpus of different sizes for training (see Table 1). This will be used to provide a clearer view of how segmentation behaves as training data becomes scarcer.

Table 1. Corpus statistics and language classification. Token counts refer to the original word-level baseline representation.

Dataset	Family/branch	Train sent.	Train tok	Dev sent.	Dev tok
Bororo 1k	Bororoan / Macro-Jê	1,000	7,856	2,138	15,678
Bororo 5k	Bororoan / Macro-Jê	5,000	37,923	2,138	15,678
Bororo full	Bororoan / Macro-Jê	17,107	128,418	2,138	15,678
Nheengatu	Tupian / Tupi-Guarani	1,077	11,811	1,043	10,002
Tupinamba	Tupian / Tupi-Guarani	464	3,593	117	915

3.3. Experimental Setup

In all experiments, we employ a BiLSTM-CRF POS tagger. The model uses a learned embedding dimension of 100, hidden dimension of 128, dropout of 0.2, AdamW optimization with learning rate 10^{-3}, batch size 16, gradient clipping at 5.0, and 10 training epochs. Each condition is trained with three random seeds: 13, 42, and 2026. For each corpus, segmentation models are trained only on the corresponding training tokens and then applied to the development set. BPE is learned with 1,000 merge operations and minimum frequency 2. Morfessor and FlatCat are trained for 10 epochs using their default settings.

The main experiments use an aggregation strategy to convert the segmented input back into a word-level sequence. Each sentence is initially represented as subtokens. For each original word, the model averages its constituent subtoken embeddings to produce a single word-level vector. The BiLSTM then processes this sequence of averaged vectors, and the CRF predicts one POS tag per word. Consequently, the loss is computed entirely at the word level. This setup maintains a standard word-level tagging task while allowing the input representations to incorporate internal subword structure.

In the following, we report and discuss the difference in word-level accuracy and macro-F1 (after removing punctuation) from the investigated methods to the baseline on the evaluation sets. Accuracy gives the overall POS tagging performance, while macro-F1 indicates whether improvements extend beyond the most frequent classes.

4. Results and Discussion

Figure 1 presents the results of our experiments, depicting the mean difference in Accuracy and Macro-F1 from the methods evaluated against the baseline. Segmentation improves accuracy in nearly every condition, with the only decrease occurring for BPE on Nheengatu. The largest gains appear in the smallest corpora: Tupinamba improves

from 0.600 to 0.684 ($+8.4pp$) with Morfessor and to 0.681 ($+8.1pp$) with BPE, while Bororo 1k improves from 0.676 to 0.732 ($+5.5pp$) with BPE. Nheengatu also shows a large gain with Morfessor, rising from 0.827 to 0.878 ($+5.1pp$). More yet, the standard deviation between the samples decreases with the size of the corpus, providing evidence that the morphological information captured by these methods improves accuracy.

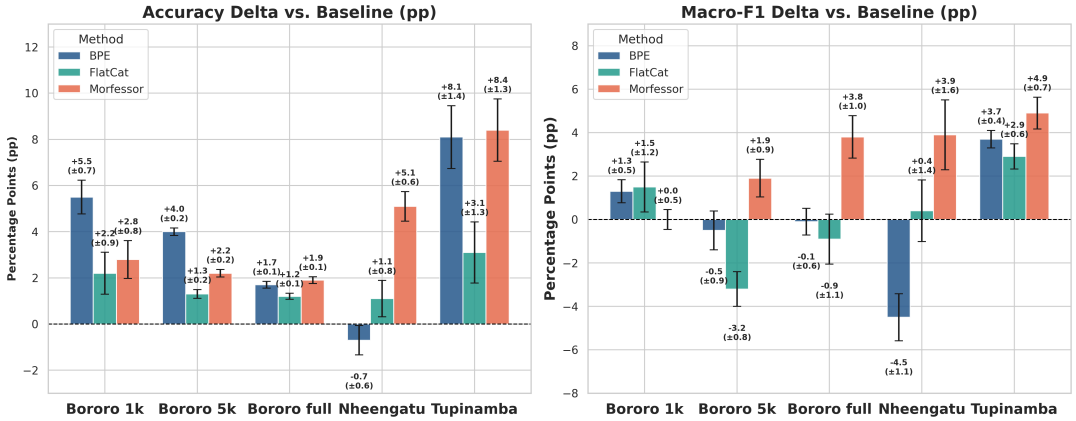


Figure 1. Mean difference (percentage points) in Accuracy and F1 from word-level baseline over three seeds, with standard deviation in error bars.

The best method varies by dataset. BPE gives the highest accuracy for the reduced Bororo settings, with gains of $+5.5pp$ in Bororo 1k and $+4.0pp$ points in Bororo 5k. Morfessor gives the highest accuracy for Bororo full, Nheengatu, and Tupinamba, and also produces the strongest macro-F1 results in most settings. FlatCat improves over the baseline in several cases, but it does not produce the best accuracy in any dataset.

These results suggest that segmentation is most useful when training data is scarce and word-level sparsity is high. The gains for Tupinamba and the Bororo data reduced settings show that subword units can help the tagger generalize across related forms. The effect, however, is not consistent across methods studied, with Morfessor being more consistent across both metrics. Notice that macro-F1 is susceptible to low performance on infrequent classes on the training data - which is often the case for low-data scenarios and a large number of classes, such as in our experiments, an aspect which we are intentionally evaluating.

These findings should be interpreted with caution. The training data scenarios explored for the Bororo language provide controlled evidence, but cross-language results also reflect differences in corpus size, annotation, domain, and morphology. Since the segmentations are evaluated only through POS tagging, the results show usefulness for tagging, not necessarily linguistically correct morphological analyses.

5. Conclusion

The results show that segmentation can improve word-level POS tagging, with Morfessor and BPE producing the strongest gains depending on the dataset and its size. Future work should examine other strategies for combining subword representations into word-level predictions, evaluate per-tag behavior in more detail, compare unsupervised segmentations against linguistically informed morphological analyses where such resources are available.

References

de Alencar, L. F. (2024). A Universal Dependencies treebank for Nheengatu. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, volume 2, pages 37–54.

Eskander, R., Lowry, C., Khandagale, S., Klavans, J., Polinsky, M., and Muresan, S. (2022). Unsupervised stem-based cross-lingual part-of-speech tagging for morphologically rich low-resource languages. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4061–4072, Seattle, United States. Association for Computational Linguistics.

Gerardi, F. F. (2023). UD_Tupinamba-TuDeT. Universal Dependencies treebank documentation.

Gerardi, F. F., Toribio, L., and Sollberger, D. (2023). UD_Bororo-BDT. Universal Dependencies treebank documentation.

Grönroos, S.-A., Virpioja, S., Smit, P., and Kurimo, M. (2014). Morfessor FlatCat: An HMM-based method for unsupervised and semi-supervised learning of morphology. In *Proceedings of COLING 2014*, pages 1177–1185.

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Keren, O., Avinari, T., Tsarfaty, R., and Levy, O. (2022). Breaking character: Are subwords good enough for MRLs after all? *CoRR*, abs/2204.04748.

Libovický, J. and Helcl, J. (2024). Lexically grounded subword segmentation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7403–7420, Miami, Florida, USA. Association for Computational Linguistics.

Mager, M., Gutierrez-Vasques, X., Sierra, G., and Meza-Ruiz, I. (2018). Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Mager, M., Oncevay, A., Mager, E., Kann, K., and Vu, T. (2022). BPE vs. morphological segmentation: A case study on machine translation of four polysynthetic languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 961–971, Dublin, Ireland. Association for Computational Linguistics.

Özçift, A., Akarsu, K., Yumuk, F., and Söylemez, C. (2021). Advancing natural language processing (nlp) applications of morphologically rich languages with bidirectional encoder representations from transformers (bert): an empirical case study for turkish. *Automatika: časopis za automatiku, mjerenje, elektroniku, računarstvo i komunikacije*, 62(2):226–238.

Saleva, J. and Lignos, C. (2021). The effectiveness of morphology-aware segmentation in low-resource neural machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 164–174, Online. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725.

Souza, F., Nogueira, R., and Lotufo, R. (2019). Portuguese named entity recognition using bert-crf. *arXiv preprint arXiv:1909.10649*.

Tonja, A., Balouchzahi, F., Butt, S., Kolesnikova, O., Ceballos, H., Gelbukh, A., and Solorio, T. (2024). Nlp progress in indigenous latin american languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6972–6987.

Tsarfaty, R., Seddah, D., Kübler, S., and Nivre, J. (2013). Parsing morphologically rich languages: Introduction to the special issue. *Computational linguistics*, 39(1):15–22.

Virpioja, S., Smit, P., Grönroos, S.-A., and Kurimo, M. (2013). Morfessor 2.0: Python implementation and extensions for Morfessor Baseline. Technical report, Aalto University.

Explorando a Integração de Múltiplas Modalidades Visuais no Reconhecimento Contínuo de Libras

Guilherme Galvão de Oliveira Pinto[1], **Guilherme Vieira Leite**[1]
Hélio Pedrini[2], **José Mario De Martino**[1]

[1]Faculdade de Engenharia Elétrica e de Computação
Universidade Estadual de Campinas

[2]Instituto de Computação
Universidade Estadual de Campinas

{g258485, g210404}@dac.unicamp.br, {pedrini, martino}@unicamp.br

Abstract. *Continuous Brazilian Sign Language (Libras) recognition presents challenges related to the multimodal nature of signing and the modeling of complex spatio-temporal dependencies. This ongoing work investigates the adaptation of neural architectures for continuous Libras recognition using RGB videos, body, facial, and hand keypoints, as well as multi-tier linguistic annotations produced in ELAN. The proposed approach combines visual feature extraction, multimodal integration, and temporal modeling. Evaluation will be conducted using the Word Error Rate (WER) metric and ablation studies. The goal is to analyze the impact of multimodality and multi-layer annotations on the performance of continuous Libras recognition models.*

Resumo. *O reconhecimento contínuo de Língua Brasileira de Sinais (Libras) apresenta desafios relacionados à natureza multimodal da sinalização e à modelagem de dependências espaço-temporais complexas. Este trabalho em andamento busca investigar a adaptação de arquiteturas neurais para o reconhecimento contínuo de Libras utilizando vídeos RGB, keypoints corporais, faciais e manuais, além de anotações linguísticas multi-tier produzidas no ELAN. O método proposto combina extração de características visuais, integração multimodal e modelagem temporal. A avaliação será realizada por meio da métrica Word Error Rate (WER) e estudos de ablação. Espera-se analisar o impacto da multimodalidade e das anotações multicamadas no desempenho de modelos de reconhecimento contínuo de Libras.*

1. Introdução

A Língua Brasileira de Sinais (Libras) constitui o principal meio de comunicação da comunidade surda brasileira e desempenha papel fundamental na promoção da acessibilidade e da inclusão social [Presidência da República 2002, Presidência da República 2015]. Diferentemente da língua portuguesa, a Libras utiliza simultaneamente componentes manuais e não manuais, incluindo configurações de mão, movimentos corporais, expressões faciais, direção do olhar e organização espacial da mensagem. Essa natureza visual-espacial e multimodal torna o desenvolvimento de tecnologias para processamento automático de línguas de sinais um desafio científico

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 594–599
October 19-22, 2026 ©2026 Association for Computational Linguistics

relevante e uma importante ferramenta para ampliar a acessibilidade em diferentes contextos sociais, educacionais e institucionais [Bragg et al. 2019].

Apesar da relevância dessas tecnologias, o reconhecimento contínuo da língua de sinais ainda é um desafio [Khan et al. 2025, Najib 2025, Wang et al. 2025]. Diferentemente de sinais isolados, reconhecê-los exige lidar com sequências de sinais de forma contínua, envolvendo fenômenos como coarticulação, transições entre sinais, variações de velocidade e a integração de múltiplos articuladores visuais. Como discutem Koller et al. [Koller et al. 2015], há uma demanda de modelos capazes de capturar dependências espaço-temporais complexas e de operar em condições mais próximas das situações reais de uso. Nos últimos anos, a literatura tem apresentado diversos métodos que buscam solucionar esses problemas. Por exemplo, Camgöz et al. [Camgoz et al. 2018] propuseram uma formulação neural para a tarefa, enquanto Camgöz et al. [Camgoz et al. 2020b] apresentaram uma arquitetura baseada em *Transformers* para reconhecimento e tradução conjuntos. Além disso, Camgöz et al. [Camgoz et al. 2020a], Yin e Read [Yin and Read 2020] e Yin et al. [Yin et al. 2023] exploraram múltiplos articuladores e novas formas de incorporar informação linguística intermediária aos modelos.

Apesar desses avanços, ainda há muitas questões em aberto, especialmente no contexto da Libras. Este trabalho está inserido em um projeto mais amplo voltado ao reconhecimento e à tradução de língua de sinais, envolvendo o desenvolvimento de um corpus multimodal de Libras contendo vídeos RGB, *keypoints* corporais, faciais e manuais, além de anotações linguísticas *multi-tier* produzidas no ELAN. O ELAN é uma ferramenta de anotação linguística que permite registrar e analisar, de forma sincronizada ao vídeo, produções em línguas de sinais [Wittenburg et al. 2006]. Este trabalho tem como objetivo principal avaliar e adaptar arquiteturas neurais para o reconhecimento contínuo, investigando o impacto de diferentes modalidades visuais e estratégias de fusão multimodal, de modo a responder à seguinte questão: como integrar, de forma consistente, vídeos RGB, *keypoints* e anotações linguísticas *multi-tier* para melhorar o desempenho de modelos neurais de reconhecimento contínuo de Libras?

2. Metodologia

A metodologia proposta visa investigar o impacto de diferentes modalidades visuais e anotações linguísticas multicamadas no reconhecimento contínuo de Libras. Três tarefas principais compõem a metodologia: construção de corpus multimodal, extração e integração de representações multimodais e modelagem temporal baseada em arquiteturas neurais.

2.1. Construção de Corpus Multimodal de Libras

Os experimentos são conduzidos com um corpus multimodal de Libras, em desenvolvimento no âmbito de um projeto de pesquisa mais amplo. O corpus é composto por sentenças em português traduzidas para Libras por uma equipe de sete intérpretes, incluindo três participantes surdos. As gravações são realizadas em ambiente controlado utilizando quatro câmeras RGB posicionadas em diferentes ângulos. Essa configuração reduz problemas de oclusão e permite a obtenção de múltiplas perspectivas da sinalização. A partir dos vídeos, são extraídas representações complementares, incluindo *keypoints*

corporais, faciais e das mãos, possibilitando a construção de diferentes modalidades de entrada para os modelos de reconhecimento.

Além das informações visuais, o corpus incorpora anotações linguísticas produzidas no ELAN [Wittenburg et al. 2006]. Diferentemente de abordagens baseadas em uma única camada de anotação, a estrutura adotada utiliza múltiplos *tiers* sincronizados temporalmente. Além das glosas, são registradas informações sobre apontamentos espaciais, mão dominante, direção do olhar e marcadores não manuais. Essa organização permite uma representação mais detalhada da sinalização e fornece informações adicionais potencialmente úteis para o treinamento de modelos neurais. Atualmente, o corpus abrange diferentes domínios de aplicação, incluindo contextos de saúde, alimentação e mobilidade, o que contribui para uma maior diversidade linguística e contextual dos dados.

2.2. Extração e Integração de Representações Multimodais

A primeira etapa do processo computacional consiste na conversão de vídeos em representações adequadas ao treinamento de modelos de reconhecimento. Os vídeos RGB são processados por extratores de características visuais, que geram *embeddings* espaciais para cada quadro da sequência. Paralelamente, representações estruturadas derivadas dos vídeos, como *keypoints* corporais, faciais e manuais, são convertidas em vetores numéricos que descrevem explicitamente a configuração espacial dos articuladores envolvidos na sinalização. Dessa forma, o reconhecimento deixa de operar diretamente sobre o vídeo bruto e passa a utilizar sequências de representações compactas que preservam informações relevantes para a modelagem da sinalização.

Um dos objetivos centrais deste trabalho é investigar como diferentes fontes de informação podem ser combinadas para apoiar o reconhecimento contínuo de Libras. Para isso, serão avaliadas diferentes estratégias de fusão multimodal. Na abordagem de fusão antecipada (*early fusion*), as representações provenientes das diferentes modalidades são combinadas antes da modelagem temporal. Já na fusão tardia (*late fusion*), cada modalidade é inicialmente processada por um fluxo independente, sendo integrada apenas em estágios posteriores da rede. A comparação entre essas estratégias permitirá avaliar em que medida as diferentes modalidades contribuem para o desempenho do sistema.

2.3. Modelagem Temporal para Reconhecimento Contínuo

O reconhecimento contínuo está sendo explorado por meio de arquiteturas neurais baseadas em *Transformers*, tomando como referência os modelos propostos por Camgöz et al. (2020) [Camgoz et al. 2020b]. A escolha dessa arquitetura se justifica por seu papel de destaque na literatura da área, tanto pelos resultados reportados quanto por oferecer uma formulação adequada para modelar sequências visuais longas, com dependências espaço-temporais complexas, características centrais do reconhecimento contínuo. A Figura 1 sintetiza as principais etapas da arquitetura, desde a obtenção das representações espaciais até a predição da sequência de glosas e a supervisão do treinamento.

Após a etapa de fusão, as sequências de representações multimodais são processadas por mecanismos de atenção capazes de modelar dependências de longo alcance ao longo da sinalização. Essa etapa busca capturar fenômenos característicos do reconhecimento contínuo, incluindo coarticulação, variações entre sinalizantes e relações contextuais entre sinais consecutivos. Como saída, o modelo produz uma sequência de glosas correspondente ao conteúdo linguístico presente no vídeo de entrada.

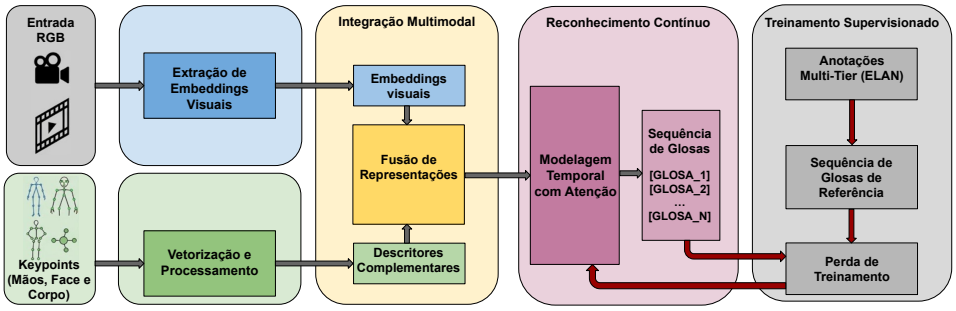


Figura 1. Fluxo computacional para o reconhecimento contínuo de Libras. As setas vermelhas indicam o uso de anotações *multi-tier* na supervisão do treinamento.

3. Avaliação Experimental

A avaliação do sistema será conduzida por meio da comparação entre as sequências de glosas produzidas pelo modelo e as anotações de referência presentes no corpus. Como métrica principal, será utilizada a *Word Error Rate* (WER), amplamente empregada em tarefas de reconhecimento contínuo de língua de sinais [Camgoz et al. 2020b]. A WER quantifica a distância de edição entre a sequência predita e a correta, considerando inserções, remoções e substituições. Além da avaliação global, serão realizados estudos de ablação com o objetivo de mensurar a contribuição individual de cada modalidade visual e das anotações multicamadas para o desempenho final do sistema.

4. Considerações Finais

Um dos principais desafios da pesquisa reside na reprodutibilidade técnica das arquiteturas de referência. A adaptação do modelo descrito em [Camgoz et al. 2020b] exige lidar com dependências de códigos, bibliotecas desatualizadas e escolhas de implementação nem sempre totalmente especificadas. Além disso, há um gargalo importante na extração de características espaciais [Alvarez et al. 2022], pois esse tipo de modelo não opera diretamente sobre o vídeo bruto, mas sobre *features* previamente extraídas. Como a arquitetura exata do extrator original não foi disponibilizada integralmente, torna-se necessário investigar alternativas para gerar *embeddings* visuais robustos, etapa central para o restante da tarefa de reconhecimento. Outro desafio diz respeito à generalização do modelo para o cenário de Libras. O sistema de referência foi desenvolvido sobre o conjunto de dados PHOENIX14T, restrito ao domínio de previsões do tempo e à Língua de Sinais Alemã, enquanto o presente trabalho se concentra na Libras e abrange domínios linguísticos mais variados.

A principal contribuição deste trabalho reside na consolidação de uma base metodológica para a adaptação e avaliação de arquiteturas neurais voltadas a esse problema, com a definição de um fluxo computacional que envolve extração de características espaciais, fusão multimodal, modelagem temporal e avaliação por WER. Embora ainda não haja resultados experimentais finais, o mapeamento dos desafios de reprodutibilidade, a extração de *features* e a generalização para o contexto de Libras já representam etapas relevantes da pesquisa. Espera-se, assim, que os próximos experimentos avancem na compreensão do impacto da multimodalidade e das anotações *multi-tier* no reconhecimento contínuo, contribuindo para o amadurecimento metodológico da área e para o desenvolvimento de tecnologias mais robustas e aplicáveis à comunidade surda brasileira.

Agradecimentos

Este trabalho é financiado pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Processo 2026/05600-6 (Bolsa de Iniciação Científica), Processo 2025/12630-6 (Bolsa de Pós-Doutorado) e Processo 2024/00914-7 (Centro de Ciência para o Desenvolvimento: Tecnologia Assistiva e Acessibilidade em Libras CCD-TAAL).

Referências

Alvarez, P. C., Nieto, X. G., and Benet, L. T. (2022). Sign Language Translation Based on Transformers for the How2Sign Dataset. *Image Processing Group Signal Theory and Communications Department Universitat Politècnica de Catalunya*.

Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreault, P., Braffort, A., Caselli, N., Huenerfauth, M., Kacorri, H., Verhoef, T., et al. (2019). Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pages 16–31.

Camgoz, N. C., Hadfield, S., Koller, O., Ney, H., and Bowden, R. (2018). Neural Sign Language Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7784–7793.

Camgoz, N. C., Koller, O., Hadfield, S., and Bowden, R. (2020a). Multi-Channel Transformers for Multi-Articulatory Sign Language Translation. In *European Conference on Computer Vision*, pages 301–319. Springer.

Camgoz, N. C., Koller, O., Hadfield, S., and Bowden, R. (2020b). Sign Language Transformers: Joint End-To-End Sign Language Recognition and Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10023–10033.

Khan, A., Jin, S., Lee, G.-H., Arzu, G. E., Dang, L. M., Nguyen, T. N., Choi, W., and Moon, H. (2025). Deep Learning Approaches for Continuous Sign Language Recognition: A Comprehensive Review. *IEEE Access*, 13:55524–55544.

Koller, O., Forster, J., and Ney, H. (2015). Continuous Sign Language Recognition: Towards Large Vocabulary Statistical Recognition Systems Handling Multiple Signers. *Computer Vision and Image Understanding*, 141:108–125.

Najib, F. M. (2025). Sign Language Interpretation using Machine Learning and Artificial Intelligence. *Neural Computing and Applications*, 37(2):841–857.

Presidência da República (2002). Lei número 10.436. Diário Oficial da União, Imprensa Nacional, n. 79, Seção 1.

Presidência da República (2015). Lei número 13.146, de 6 de julho de 2015. Institui a Lei Brasileira de Inclusão da Pessoa com Deficiência (Estatuto da Pessoa com Deficiência). Diário Oficial da União, Brasília, DF.

Wang, Z., Li, D., Jiang, R., and Okumura, M. (2025). Continuous Sign Language Recognition with Multi-Scale Spatial-Temporal Feature Enhancement. *IEEE Access*, 13:5491–5506.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). ELAN: A Professional Framework for Multimodality Research. In *5th International Conference on Language Resources and Evaluation*, pages 1556–1559.

Yin, A., Zhong, T., Tang, L., Jin, W., Jin, T., and Zhao, Z. (2023). Gloss Attention for Gloss-Free Sign Language Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern rRecognition*, pages 2551–2562.

Yin, K. and Read, J. (2020). Better Sign Language Translation with STMC-Transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5975–5989.

Da Definição das Entidades à Anotação do Corpus: Desafios da tarefa no Direito do Consumidor

Samuel Rios da Silva[1], **Aline Athaydes**[1], **Babacar Mane**[1] , **Daniela Barreiro Claro**[1]

[1]FORMAS Research Center on Data and Natural Language
Institute of Computing – Federal University of Bahia (UFBA)
Av. Milton Santos – Campus de Ondina – 40.170-110 – Salvador – BA – Brazil

{samuelrs,alineathaydes,mbabacar,dclaro}@ufba.br

Abstract. *Data annotation is a crucial step in the development of machine learning models aimed at Named Entity Recognition (NER). In this context, given the identified need and the lack of specialized NER models for the Brazilian Consumer Defense Code (CDC), this work presents the process of annotating a corpus with specific entities related to the CDC, as well as the challenges encountered, the lessons learned, the main results obtained, and our next steps.*

Resumo. *A anotação de dados é uma etapa crucial para a construção de modelos de apredizado de máquina voltados ao Reconhecimento de Entidades Nomeadas (REN). Nesse sentido, percebida a necessidade e a ausência de Modelos REN especialistas no contexto do Código de Defesa do Consumidor (CDC), este trabalho apresenta o processo de anotação de um corpus com entidades específicas relacionadas ao CDC, assim como os desafios encontrados, os aprendizados, principais resultados obtidos e nossos próximos passos.*

1. Introdução

Os modelos de Reconhecimento de Entidades Nomeadas (REN) são capazes de identificar e classificar entidades mencionadas em um grande volume de dados textuais, facilitando sua análise semântica. Nesse contexto, a anotação de um *corpus* para o treinamento desses modelos é uma atividade essencial de interpretação e classificação humana e/ou automática, que leva em consideração o contexto linguístico [Caseli and Nunes 2026].

No âmbito do Código de Defesa do Consumidor (CDC) [Brasil 1990], um tópico específico do domínio jurídico, a literatura lista um conjunto de corpora voltados para o REN jurídico [Caseli and Nunes 2026]. Em sua maioria, são compostos por documentos formais e possuem entidades genéricas ou específicas de outros tópicos, não abrangendo entidades representativas adequadas para o CDC.

Sendo o CDC uma referência central para os direitos do consumidor [Athaydes et al. 2024, Athaydes et al. 2025], identificamos uma lacuna na geração de um conjunto de etiquetas especializadas nesse código, focado na entrada de textos redigidos em linguagem coloquial por usuários. Este trabalho, portanto, apresenta o processo de criação dessas etiquetas e anotação de um *corpus* com entidades relacionadas ao CDC, destacando os desafios encontrados, os resultados obtidos e as lições aprendidas ao longo do processo.

Este artigo está estruturado como segue: a Seção 2 apresenta os trabalhos relacionados, a Seção 3 descreve a construção do *corpus*. As Seções 4 e 5 apresentam o

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 600–604
October 19-22, 2026 ©2026 Association for Computational Linguistics

conjunto de entidades e a ferramenta de anotação respectivamente. A Seção 6 descreve os resultados e a conclusão é apresentada na Seção 7.

2. Trabalhos Relacionados

Nos últimos anos, trabalhos avançaram sobre documentos legislativos brasileiros. O UlyssesNER-Br ampliou os recursos para PLN jurídico ao explorar textos legislativos como fonte para anotação de entidades [Albuquerque et al. 2022]. Nunes et al. [Nunes et al. 2024] também investigaram REN em textos legislativos em português com autoaprendizado. Esses estudos indicam que o REN jurídico em português tem se concentrado principalmente em documentos formais e institucionais.

Apesar desses avanços, ainda há uma lacuna quando o foco se desloca de documentos formais para relatos de consumidores escritos em linguagem coloquial. No contexto do Código de Defesa do Consumidor (CDC), entidades como consumidor, fornecedor, produto, serviço e relação de consumo podem aparecer de forma implícita, incompleta ou ambígua, exigindo critérios de anotação específicos. Assim, este trabalho propõe uma definição de etiquetas especializadas e apresenta os desafios de anotação dos relatos dos consumidores, aplicando a tarefa de REN jurídico a textos produzidos por usuários finais no domínio do CDC.

3. Construção do Corpus

Textos jurídicos, assim como textos do CDC, contêm frequentemente linguagem técnica e sentenças longas [Bessa et al. 2026], dificultando a compreensão por grande parte de seus leitores. Com o intuito de obter um conjunto de dados mais próximos aos descritos pelos usuários, optou-se pela plataforma *Consumidor.gov.br*, que reúne relatos (ou reclamações) redigidos na linguagem do usuário final. O *corpus* foi obtido por meio da extração de mil relatos de consumidores com as suas respectivas relações de consumo.

4. Conjunto de Entidades

As entidades a serem anotadas foram definidas e extraídas diretamente dos artigos 2º e 3º do CDC [Brasil 1990]:

- ***Consumidor***: Toda pessoa física ou jurídica que adquire ou utiliza produto ou serviço como destinatário final.
- ***Fornecedor***: Toda pessoa física ou jurídica, pública ou privada, nacional ou estrangeira, bem como os entes despersonalizados, que desenvolvem atividade de produção, montagem, criação, construção, transformação, importação, exportação, distribuição ou comercialização de produtos ou prestação de serviços.
- ***Produto***: Qualquer bem, móvel ou imóvel, material ou imaterial.
- ***Serviço***: Qualquer atividade fornecida no mercado de consumo, mediante remuneração, inclusive as de natureza bancária, financeira, de crédito e securitária, salvo as decorrentes das relações de caráter trabalhista.

Com essas definições iniciais estabelecidas, foi realizada uma primeira reunião com uma especialista no domínio jurídico para validação desse conjunto de entidades. Embora as categorias Consumidor, Fornecedor, Produto e Serviço estejam definidas nos artigos 2º e 3º do CDC, observou-se que, nos relatos analisados, a identificação do consumidor muitas vezes é implícita e está associado à relação estabelecida entre o usuário

e o fornecedor. Como resultado, observou-se a necessidade de identificar os eventos que estabelecem a relação de consumo. Foi, então, adicionada o evento *RelacaoConsumo*, definida como segue:

- **RelacaoConsumo**: Um evento que estabelece uma relação de consumo, por exemplo: ”comprei”ou ”contratei”.

O acrécimo de tal evento, estende a tarefa de REN para incluir o reconhecimento de Eventos [Caseli and Nunes 2024].

5. Ferramenta de Anotação

Um total de 50 textos aleatórios foram selecionados para anotação manual e validação por duas especialistas. As anotações ocorreram através da ferramenta Label Studio [Tkachenko et al. 2025], uma plataforma de código aberto que permite selecionar trechos do texto e associá-los às entidades. A Figura 1 exibe sua interface já com o conjunto de etiquetas definidos para essa tarefa.

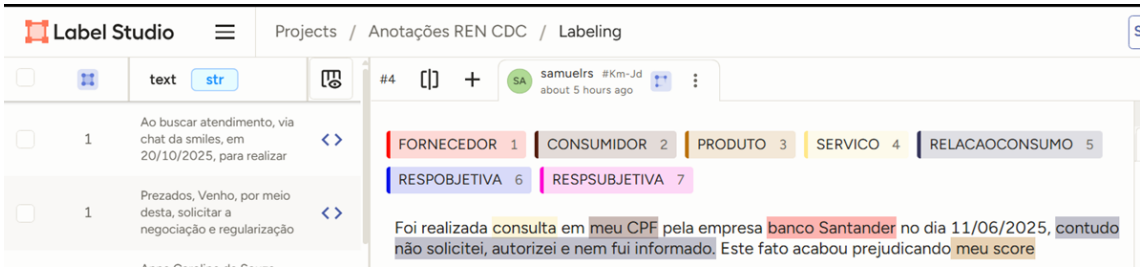


Figura 1. Interface do Label Studio para anotação dos relatos do Consumidor.gov.br

6. Resultados e Discussões

Para avaliar a concordância entre as anotadoras, adotaram-se duas métricas: o Kappa de Cohen (κ) adaptado para janelas de sobreposição (spans) e o F1-Score Cruzado (Pairwise F1). O resultado para o Kappa de Cohen foi de $\kappa = 0,1074$ (10%) e para o F1-Score Cruzado obtivemos um resultado de $0,3088$ (30%). Tais resultados apontam para uma baixa concordância geral entre as anotadoras.

Após a anotação, uma discussão com as especialistas buscou mapear os motivos dessa divergência que seguem descritos. O primeiro problema observado foi a falta de um **manual de anotação (*guideline*)**, fazendo com que as especialistas concluíssem que era obrigatória a anotação das entidades *Consumidor*, *Fornecedor* e *Produto*. Esse entendimento equivocado por parte das especialistas favoreceu com que cada uma adotasse **convenções próprias** para identificar tais entidades que não estavam presente explicitamente no relato (por exemplo, “meu CPF” era marcado como *Consumidor*). Além disso, houve **interpretações divergentes** sobre a entidade *RelacaoConsumo*, e **trocas frequentes entre os conceitos** de *Produto* e *Serviço* entre ambas as especialistas.

A partir dessa etapa, observou-se a necessidade de um alinhamento prévio em conjunto com ambas as especialistas. Uma vez alinhado em um segundo momento e as dúvidas sanadas, criou-se um *Manual de Anotação* formalizando as regras que deveriam

ser seguidas de acordo com o CDC. A última versão desse manual, abrange as seguintes entidades: *Consumidor*, *Fornecedor, Produto, Serviço, RelaçãoConsumo*. O manual está em fase final de refinamento para aprimorar essas definições e solucionar discussões em aberto.

Tabela 1. Resultado das Métricas por Entidade

Entidade	Kappa de Cohen	F1-Score Cruzado	Instâncias (Anotadora 1)	Instâncias (Anotadora 2)
CONSUMIDOR	0,4728	0,5591	44	49
FORNECEDOR	0,7764	0,8095	41	43
PRODUTO	0,0368	0,0435	1	45
RELACAOCONSUMO	-0,1140	0,0808	49	50
SERVICO	0,1132	0,2581	49	50

A partir do manual e o alinhamento entre as especialistas, as concordâncias entre as anotadoras foram estratificadas por etiquetas (Tabela 1), e observou-se um aumento significativo nas entidades *Fornecedor* e *Consumidor*. A entidade Serviço ainda carece de análise assim como a entidade *Produto*. A pior discordância foi observada na *RelaçãoConsumo*, visto que não há um acordo entre as especialistas, sendo necessária uma nova rodada de alinhamento ou a retirada dessa etiqueta do processo de anotação. Assim, o presente trabalho apresentou os desafio de uma tarefa de anotação, mesmo quando se tem a presença de especialistas no domínio.

7. Conclusões e Trabalhos Futuros

A análise preliminar revelou baixa concordância inicial, mas o F1-Score de 80% para Fornecedor demonstra o potencial da tarefa. Ficou claro que o maior desafio ao transpor conceitos do CDC para a linguagem coloquial é definir um conjunto representativo de entidades com precisão.

Essa baixa concordância derivou da falta de um manual e de alinhamento prévio, gerando interpretações subjetivas e confusões (como entre Produto e Serviço). Isso evidencia a necessidade de ciclos iterativos; o alinhamento posterior com especialistas para criar o Manual de Anotação foi essencial para padronizar o processo.

Como trabalhos futuros, prevê-se um aumento significativo na concordância entre anotadoras para consolidar um *corpus* ouro (*gold standard*). A partir dele, planeja-se a anotação automática dos dados restantes via Grandes Modelos de Linguagem (LLMs), com revisão humana para garantir confiabilidade e agilizar a construção do *corpus* especialista no CDC.

Agradecimentos

Os autores agradecem o financiamento parcial através da bolsa FAPESB 2025/2026, do apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001, FAPESB CCE 0022/2023, FAPESB TIC 002/2015. Este trabalho também conta com o apoio do Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação (INCT-TILD-IAR) (processo 408490/2024-1). Gostaríamos de agradecer ao Escavador pela parceria neste trabalho.

Referências

Albuquerque, H. O., Costa, R., Silvestre, G., Souza, E., da Silva, N. F. F., Vitório, D., Moriyama, G., Martins, L., Soezima, L., Nunes, A., Siqueira, F., Tarrega, J. a. P., Beinotti, J. V., Dias, M., Silva, M., Gardini, M., Silva, V., de Carvalho, A. C. P. L. F., and Oliveira, A. L. I. (2022). Ulyssesner-br: A corpus of brazilian legislative documents for named entity recognition. In *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings*, page 3–14, Berlin, Heidelberg. Springer-Verlag.

Athaydes, A., Bulcao, L., Sacramento, C., Mane, B., Claro, D. B., Souza, M., and Pita, R. (2024). Brazilian consumer protection code: a methodology for a dataset to question-answer (QA) models. In Claro, D. B. and Pagano, A., editors, *Proceedings of the 15th Brazilian Symposium in Information and Human Language Technology*, pages 522–529, Belém do Pará, Brazil. Association for Computational Linguistics.

Athaydes, A., Mota, L., Neto, F. M., Silva, S., Mane, B., Claro, D., Souza, M., and Lisboa, A. (2025). Towards a corpus methodology for llms in the legal domain. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 272–282, Porto Alegre, RS, Brasil. SBC.

Bessa, A. C. C., Lobato, F. M. F., and Junior, A. F. L. J. (2026). Development and evaluation of a hybrid information retrieval system applied to the Brazilian legal domain. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 2*, pages 186–190, Salvador, Brazil. Association for Computational Linguistics.

Brasil (1990). Lei nº 8.078, de 11 de setembro de 1990. Dispõe sobre a proteção do consumidor e dá outras providências. Diário Oficial da União, Brasília, DF.

Caseli, H. d. M. and Nunes, M. d. G. V. (2024). Extração de informação. In Caseli, H. d. M. and Nunes, M. d. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações*, chapter Extração de Informação. 3º edition.

Caseli, H. M. and Nunes, M. G. V., editors (2026). *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, volume 3. BPLN, 4 edition.

Nunes, R. O., Balreira, D. G., Spritzer, A. S., and Freitas, C. M. D. S. (2024). A named entity recognition approach for Portuguese legislative texts using self-learning. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 290–300, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Tkachenko, M., Malyuk, M., Holmanyuk, A., and Liubimov, N. (2020-2025). Label Studio: Data labeling software. Open source software available from https://github.com/HumanSignal/label-studio. Acesso em: 02 jun. 2026.

Anotação de Metáforas em Saúde Mental: Análise de Postagens Baseada no Procedimento MIPVU

Isabela C. Rodrigues[1], **Helena de Medeiros Caseli**[2]

[1]Departamento de Letras (DL) — Universidade Federal de São Carlos (UFSCar)

[2]Departamento de Computação (DC) — Universidade Federal de São Carlos (UFSCar)
Rod. Washington Luís, km 235 — 13565-905 — São Carlos — SP — Brasil

isabelarodrigues@estudante.ufscar.br, helenacaseli@ufscar.br

Abstract. *This paper presents the preliminary results of a systematic investigation into the use of metaphors related to depression in social media posts in Brazilian Portuguese. Based on the MIPVU annotation procedure and Cognitive Linguistics, 743 metaphorical expressions were identified across 13 conceptual domains. The results highlight the centrality of the Journey and Delimited Space/Container domains, as well as the importance of differentiating between First-Person (1P) and Other-Person (OP) perspectives to capture subjective experiences and avoid false positives in NLP models. The study also proposes emerging domains such as Quantity/Size and Sensory Perception.*

Resumo. *Este trabalho apresenta os resultados preliminares de uma investigação sistemática sobre o uso de metáforas em postagens de redes sociais relacionadas à depressão em português do Brasil. Com base no procedimento de anotação MIPVU e na Linguística Cognitiva, foram identificadas 743 expressões metafóricas distribuídas em 13 domínios conceituais. Os resultados destacam a centralidade dos domínios de Jornada e Espaço Delimitado/Recipiente, além da importância da diferenciação entre as perspectivas de Primeira Pessoa (1P) e Outra Pessoa (OP) para capturar a vivência subjetiva e evitar falsos positivos em modelos de PLN. O estudo também propõe domínios emergentes como Quantidade/Tamanho e Percepção Sensorial.*

1. Introdução

A depressão constitui um transtorno de saúde mental que afeta uma parcela significativa da população, sendo caracterizada por um sofrimento psicológico clinicamente significativo que compromete a funcionalidade, reduz a qualidade de vida e, em casos graves, eleva o risco de morte por suicídio. Estima-se que apenas um quarto dos indivíduos com transtornos mentais recebam cuidados adequados [Organização Mundial da Saúde 2021], o que tem impulsionado pesquisas em Processamento de Linguagem Natural (PLN) focadas na detecção e intervenção precoce [Paraboni and Caseli 2026].

Nas redes sociais, os usuários frequentemente recorrem a recursos simbólicos para expressar emoções subjetivas e complexas, dos quais destaca-se a **metáfora** que opera não apenas como ornamento estilístico, mas como uma estrutura essencial do pensamento humano. Na Linguística Cognitiva, o sistema conceitual humano é em grande parte metafórico, permitindo que domínios abstratos e complexos sejam compreendidos em termos

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 605–609
October 19-22, 2026 ©2026 Association for Computational Linguistics

de conceitos mais concretos e acessíveis [Lakoff and Johnson 1980]. Em desabafos emocionais, as metáforas conferem expressividade e afetividade, permitindo a expressão de experiências subjetivas em frases de alto impacto [Shi and Khoo 2023].

Nesse contexto, técnicas de PLN têm sido usadas para análises postagens online voltadas à saúde mental. Na dimensão figurativa, [Shi and Khoo 2023] destaca que a análise de metáforas em comunidades virtuais revela como indivíduos externalizam o sofrimento psíquico por meio de domínios conceituais específicos, corroborando a premissa de que a metáfora não é apenas um recurso estilístico, mas uma ferramenta cognitiva essencial para dar sentido à experiência subjetiva da depressão [Roystonn et al. 2021]. Essas experiências são descritas com metáforas de "descida", "escuridão", "peso físico" ou "aprisionamento", conforme evidenciado em relatos de primeira pessoa em blogs [Coll-Florit et al. 2021]. No entanto, a maioria dos trabalhos que analisam o uso de linguagem figurada na depressão foca no idioma inglês, evidenciando a lacuna em português.

Diante disso, o presente trabalho apresenta resultados preliminares sobre o uso de metáforas em postagens de redes sociais em português utilizando o procedimento de anotação sistemático MIPVU. 200 postagens foram analisadas resultando em 743 anotações de unidades metafóricas em treze domínios relacionados à saúde mental. Além das anotações, outras contribuições deste trabalho são a proposta de considerar a perspectiva enunciativa (primeira pessoa ou outra pessoa) na anotação e quatro novos domínios.

2. Metodologia

Os dados analisados integram o *corpus* Amive, composto por 604 postagens do Facebook em português do Brasil anotadas com 18 sintomas de depressão [Wilkens et al. 2026]. Neste estudo, foram analisadas 200 postagens na plataforma INCEpTION[1] com a definição de uma camada específica para a identificação de unidades metafóricas e seus domínios e perspectivas enunciativas. O processo de anotação fundamenta-se no MIPVU (*Metaphor Identification Procedure Vrije Universiteit*) [Steen et al. 2010] e a aplicação do método seguiu um fluxo estruturado em cinco etapas:

1. Leitura integral do texto para estabelecer um entendimento sobre a intenção comunicativa, especialmente em relação ao tema da saúde mental.
2. Identificação das unidades lexicais (palavra ou expressão mínima com significado próprio) que abrange desde vocábulos isolados até estruturas complexas, como substantivos compostos e expressões idiomáticas.
3. Verificação da metaforicidade por meio do contraste semântico onde o sentido da unidade no contexto é confrontado com um significado contemporâneo mais básico[2]. Se o sentido básico contrastar com o contextual, mas puder ser compreendido em comparação a ele, a unidade é marcada como metáfora.
4. Atribuição de etiquetas que distinguem a perspectiva enunciativa entre Primeira Pessoa (1P) e Outra Pessoa (OP). A etiqueta 1P é empregada para capturar a vivência subjetiva e a autoexpressão da depressão, enquanto a etiqueta OP mapeia a percepção do sujeito sobre o ambiente externo, englobando julgamentos de terceiros, estigmas sociais ou a relação com outras pessoas. Essa distinção é essencial

[1] `https://inception-project.github.io/`

[2] Segundo [Steen et al. 2010], o sentido mais básico ou mais concreto é aquele relacionado ao corpo ou ao movimento físico. Para essa verificação, utiliza-se o Dicionário Houaiss como suporte lexicográfico.

assegurando que metáforas relacionadas a terceiros não sejam erroneamente atribuídas ao autor da postagem como indicadores de sintomas.

5. Categorização das metáforas com base em domínios concretos e abstratos validados por estudos qualitativos e linguísticos que exploram as vivências subjetivas da depressão [Lakoff and Johnson 1980, Coll-Florit et al. 2021, Roystonn et al. 2021]. Os eixos temáticos de maior relevância e recorrência na literatura são a "Depressão como Aflição", manifestada primordialmente por metáforas de (1) escuridão (**Luz/Escuridão**) e de (2) confinamento em espaços restritos (**Espaço delimitado/Recipiente**), e a "Depressão como Jornada", que estrutura a experiência do transtorno como um (3) percurso espacial ou movimento (**Jornada**), podendo indicar (4) progresso ou retrocesso (**Para cima/Para baixo**) [Coll-Florit et al. 2021, Roystonn et al. 2021]. Estão presentes também os domínios (5) **Inimigo/Guerra** e (6) **Força**, que sistematizam a vivência como uma batalha árdua ou uma pressão externa sobre o indivíduo; e (7) **Fardo/Peso** e (8) **Organismo vivo**, que personificam sentimentos como entidades autônomas [Coll-Florit et al. 2021]. Por fim, a literatura aponta o domínio (9) **Coisa/Incômodo**, que conceitualiza o transtorno e o sofrimento como objetos físicos, permitindo ações de descarte ou afastamento de estados negativos, bem como a oferta/recebimento de suporte. A aplicação prática das diretrizes e a recorrência de novos padrões no *corpus* levaram à identificação de domínios adicionais: (10) **Céu/Inferno**, que descreve estados de sofrimento extremo ou alívio espiritual; (11) **Máquina**, que captura a perda de vitalidade e o funcionamento do sujeito de forma mecânica; (12) **Percepção sensorial**, que recruta os sentidos biológicos para descrever processos cognitivos; e (13) **Quantidade/Tamanho**, que traduz a intensidade emocional em dimensões físicas. Assim, no total 13 domínios diferentes foram considerados e a anotação foi realizada, inclusive, com a possibilidade de atribuição de mais de um domínio à mesma unidade.

As 200 postagens foram anotadas manualmente por uma única anotadora (primeira autora deste artigo), graduanda em Linguística e bolsista de Iniciação Científica, sob supervisão da pesquisadora responsável (segunda autora). A anotação compreendeu dois meses e um mês seguinte foi dedicado à revisão dos dados, à análise dos resultados e ao refinamento das diretrizes para as primeiras 100 postagens. O processo se repetiu por mais 2 meses para as outras 100 postagens.[3]

Para ilustrar a anotação e preservar a privacidade, a seguir utilizam-se exemplos fictícios, sem reproduzir postagens do *corpus*. Em "Sinto-me *presa* nessa tristeza", *presa* contrasta com o sentido básico de confinamento físico e recebe as etiquetas Espaço Delimitado/Recipiente e 1P. Em "Eles não *enxergam* minha dor", *enxergam* contrasta com o sentido básico de percepção visual e recebe as etiquetas Percepção Sensorial e OP.

3. Resultados

No total foram identificadas 743 expressões metafóricas no *corpus* analisado. Os dados revelam uma diversidade significativa de domínios-fonte utilizados para externalizar a experiência da depressão, distribuídos conforme a Tabela 1.

[3]A camada de anotação metafórica não contou com a participação direta de especialistas em saúde mental e também não foi possível calcular uma medida de concordância interanotador, como o coeficiente Kappa, porque até o momento a anotação foi realizada por apenas uma pessoa.

Tabela 1. Consolidação dos Domínios Metafóricos

Domínio	1P	OP	Total	Domínio	1P	OP	Total
ESPAÇO DELIMITADO/RECIPIENTE	90	43	133	QUANTIDADE/TAMANHO	20	12	32
JORNADA	82	36	118	PERCEPÇÃO SENSORIAL	12	22	34
COISA/INCÔMODO	54	33	87	FARDO/PESO	38	05	43
FORÇA	56	31	87	LUZ/ESCURIDÃO	11	06	17
INIMIGO/GUERRA	39	23	62	CÉU/INFERNO	04	03	07
ORGANISMO VIVO	28	20	48	MÁQUINA	04	03	07
PARA CIMA/PARA BAIXO	23	15	38	OUTRO	–	–	30

Domínios Predominantes – A experiência depressiva no *corpus* é predominantemente estruturada por metáforas de espacialidade e movimento. O domínio **Espaço Delimitado/Recipiente** foi o mais frequente, com 133 ocorrências, evidenciando a tendência de conceituar o transtorno como um estado de isolamento ou aprisionamento, com forte incidência em primeira pessoa (90). O domínio **Jornada**, com 118 ocorrências, foi amplamente utilizado para descrever a trajetória temporal da doença e o processo terapêutico. Os domínios **Coisa/Incômodo** e **Força** apresentaram 87 ocorrências cada. No primeiro, estados mentais são tratados como objetos físicos passíveis de serem carregados ou descartados; no segundo, o sofrimento é relatado como uma pressão externa que limita a ação do indivíduo.

Perspectivização e Alteridade – Observa-se uma concentração maior de metáforas em 1P (461) em relação a OP (252). Essa tendência se mantém em quase todos os domínios com exceção do **Percepção Sensorial** (34 ocorrências), que se destaca pela predominância em OP (22 ocorrências). Essa distribuição indica que os autores utilizam falhas sensoriais para representar a falta de escuta ou de compreensão por parte de terceiros, ilustrando simbolicamente a ruptura na comunicação.

Domínios Emergentes e Nuances do *Corpus* – Em relação aos quatro novos domínios propostos, **Percepção Sensorial** registrou 34 ocorrências e destacou-se pela predominância em OP, conforme discutido anteriormente. **Quantidade/Tamanho** (32) revelou-se um recurso importante para traduzir a intensidade emocional em dimensões físicas e volume, conferindo uma magnitude tangível à dor subjetiva; **Céu/Inferno** (07) é empregado para descrever estados extremos, oscilando entre o sofrimento agudo e o alívio espiritual ou punição; e **Máquina** (07) captura a sensação de perda de vitalidade e o funcionamento do sujeito de forma mecânica ou em "piloto automático", indicando um distanciamento da própria agência humana.

4. Conclusões

Este estudo identificou 743 expressões metafóricas em 200 postagens, destacando a recorrência dos domínios **Espaço Delimitado/Recipiente** e **Jornada** e a contribuição da distinção entre 1P e OP. Como limitação, por se tratar de um estudo preliminar, as anotações foram realizadas por uma única pessoa, sem cálculo de concordância interanotador.

Como trabalhos futuros, pretende-se ampliar o *corpus* anotado e investigar a viabilidade de um processo semiautomático baseado no paradigma *Human-in-the-Loop* onde um modelo computacional adaptado ao português do Brasil produzirá pré-anotações de possíveis unidades metafóricas, que serão importadas na INCEpTION e submetidas à curadoria humana que fará a associação aos domínios. Essa abordagem permitirá comparar o tempo e o esforço da curadoria assistida com os da anotação inteiramente manual.

Agradecimentos

O presente trabalho foi realizado com apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Brasil. Processos nº 2025/09978-0 (IC) e nº 2024/10233-7 (Auxílio Regular, projeto AIM-Health). Em conformidade com as diretrizes do evento, as autoras declaram o uso do modelo generativo de IA Gemini (Google) para auxiliar na formatação em LaTeX. O conteúdo, as análises e as conclusões são responsabilidade integral dos autores.

Referências

Coll-Florit, M., Climent, S., Sanfilippo, M., and Hernández-Encuentra, E. (2021). Metaphors of depression. studying first person accounts of life with depression published in blogs. *Metaphor and Symbol*, 36(1):1–19.

Lakoff, G. and Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press, Chicago and London.

Organização Mundial da Saúde (2021). https://www.who.int/pt/news-room/fact-sheets/detail/depression. Acesso em: 4 maio 2026.

Paraboni, I. and Caseli, H. d. M. (2026). Detecção de transtornos de saúde mental a partir de texto. In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, volume 3, book chapter 10. BPLN, 4 edition.

Roystonn, K., Teh, W. L., Samari, E., Cetty, L., Devi, F., Shahwan, S., Chandwani, N., and Subramaniam, M. (2021). Analysis and interpretation of metaphors: Exploring young adults' subjective experiences with depression. *Qualitative Health Research*, 31(8):1437–1447.

Shi, J. and Khoo, Z. (2023). Words for the hearts: a corpus study of metaphors in online depression communities. *Frontiers in Psychology*, 14:1227123.

Steen, G. J., Dorst, A. G., Herrmann, J. B., Kaal, A., Krennmayr, T., and Pasma, T. (2010). *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Wilkens, R., Caseli, H., Neris, V., and Villavicencio, A. (2026). The visible and the latent linguistic clues of mental health in Brazilian Portuguese textual posts. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 2*, pages 135–147, Salvador, Brazil. Association for Computational Linguistics.

Esquecida no Churrasco: Um método para validação e correção responsável de traduções automáticas do BBQ para Português do Brasil

Luiza Rodrigues Cardoso, Leo Sampaio Ferraz Ribeiro

[1]USP, Instituto de Ciências Matemáticas e Computação - São Carlos, SP – Brasil

luizarodriguescardoso@usp.br, leo.ribeiro@icmc.usp.br

Abstract. *Com os rápidos avanços em Processamento de Linguagem Natural, torna-se necessário o desenvolvimento de ferramentas de avaliação de performances dos modelos linguísticos diante da destilação de vieses sociais. Neste artigo, então, apresentamos um método de adaptação de datasets que contempla adequações de fatores linguísticos e socioculturais do Brasil, uma demonstração de aplicação do método ao dataset BBQ e a descrição da implementação da nossa plataforma autoral de validação.*

1. Introdução

Com o avanço de modelos de linguagem (LMs) e suas aplicações abrangentes, surge a preocupação acerca das consequências sociotécnicas, como a disseminação de vieses nos conteúdos gerados por estes modelos, como os utilizados em chatbots [Bonil et al. 2025].

O termo viés, em Aprendizado de Máquina, é usado para descrever os tipos de modelos que podem ser obtidos via diferentes algoritmos e as limitações deste espaço de busca. Neste artigo, contudo, usamos o termo sob o aspecto ético, em que viés descreve, então, os impactos dos estereótipos sociais sobre os dados, fatores humanos que, quando agregados aos dados, desbalanceiam os modelos nos treinamentos por meio da sub ou super-representação de conceitos [Bender et al. 2021]. Este tipo de aprendizado resulta em comportamentos reprodutores de visões hegemônicas [Messeri and Crockett 2024] e associações enviesadas sobre gênero, raça e outras características de identidade pessoal, bem como padrões sistêmicos de discriminação [Talat et al. 2022]. Esse problema pode ser visto especialmente em PLN, em que os dados de treinamento – *corpus* – são obtidos por meio de fóruns online ou obras literárias em domínio público e, portanto, refletem opiniões pessoais, incluindo estereótipos, que afetam os desempenhos dos modelos frente aos padrões de ética e moral.

Com a adoção generalizada destes modelos, são necessárias garantias tanto de desempenho quanto de segurança. Segundo [Neplenbroek et al. 2024], há uma diferença de desempenho de LLMs em *prompts* que utilizam idiomas não anglófonos, ocorrendo um agravamento da disseminação de vieses sociais. Possivelmente esse baixo desempenho decorre do fato de que o idioma dominante durante a fase de treinamento – chegando a 90% do total – e utilizado na etapa de treinamento de segurança é o inglês [OpenAI et al. 2024, Touvron et al. 2023, Brown et al. 2020].

O trabalho desenvolvido por [Parrish et al. 2022] traz uma notável contribuição na tarefa de avaliação da performance de LLMs frente aos vieses sociais, em que, percorrendo 9 categorias sociais baseadas no aspecto sociocultural dos EUA, e utilizando

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 610–615
October 19-22, 2026 ©2026 Association for Computational Linguistics

templates, com ajustes de agentes e objetos das sentenças, foi possível identificar que os modelos reproduzem estereótipos de acordo com os parâmetros e nível de descrição dos agentes das sentenças.

Assim, surge a necessidade de investigar o comportamento destes modelos em outros contextos sociais e linguísticos, como os do Brasil. Neste trabalho, então, propomos (1) Uma plataforma para validação e submissão de *templates*; (2) Um método de validação de traduções diretas de datasets, em que, a partir da plataforma, o avaliador poderá aprovar a tradução direta e sugerir alterações; (3) Uma versão inicial do dataset BBQ em português brasileiro.

2. Trabalhos Relacionados

Como [Bender et al. 2021, Talat et al. 2022] demonstram em seus trabalhos, os modelos possuem a tendência de amplificar discursos prejudiciais da sociedade, sendo necessário o desenvolvimento de ferramentas que mitiguem esses comportamentos. Há iniciativas que buscam compreender e avaliar a presença de vieses relacionados ao gênero, identidade não-binária, religião, raça e temas relacionados a pessoas com deficiências [Ovalle et al. 2023, Zhao et al. 2021, Gadiraju et al. 2023], porém a maioria dos trabalhos é focada na língua inglesa.

Neste sentido diversos benchmarks de auditoria foram criados, como o CrowS-pairs [Nangia et al. 2020], StereoSet [Nadeem et al. 2020] e BBQ [Parrish et al. 2022]. Destacamos este último como um caso de estudo para esse trabalho; o BBQ foi criado para avaliar para modelos Question-Answering que lida com múltiplos aspectos sociais, baseando-se em situações que servem para estudar os níveis de vieses sociais, tendo como limitação os aspectos socioculturais dos Estados Unidos. O trabalho utiliza *templates*, com os parâmetros de agente e objetos adaptáveis, que permitem construir os cenários destes agentes e permutar diferentes estereótipos presentes na sociedade estudada. Diante disso, surgiram algumas variantes do BBQ, o Korean-BBQ [Jin et al. 2024], Chinese-BBQ [Huang and Xiong 2023], German Gender BBQ [Satheesh et al. 2025] e o Multilingual-BBQ [Neplenbroek et al. 2024]. Evidencia-se a adaptabilidade deste dataset para outros idiomas e contextos socioculturais. Assim, propomos uma versão brasileira com validação humana e adaptativa dos *templates* ao nosso contexto sociocultural.

Outros datasets que visam realizar este tipo de avaliação sob o aspecto regional são SHADES [Mitchell et al. 2025] e LACES [Ivetta et al. 2025]. Sendo o primeiro desenvolvido para avaliar estereótipos culturais e específicos presentes em modelos linguísticos sob 37 regiões do globo. E o segundo possui enfoque na região latino-americana e uma coleta participativa, ou seja, a comunidade possui participação na construção, gerando um conjunto realista. Ambos agregam à tarefa de avaliação estereótipos específicos e nuances culturais. Esses aspectos, em consonância, compõem a nossa proposta.

No Brasil, há uma crescente movimentação para estudar a proliferação dos vieses sob os aspectos do nosso idioma e cultura [Silva et al. 2024]. Alguns trabalhos se focam na identificação da questão de gênero na tarefa de tradução automática [Soares et al. 2023] e na redução de vieses sociais em modelos baseados na arquitetura CLIP, o FairPivara proposto por [Moreira et al. 2024]. Outro trabalho fundamental é o desenvolvido por [Bonil et al. 2025], que avalia as questões de gênero e raça em narrativas geradas pelos LLMs,

Esses trabalhos evidenciam os diferentes tipos de abordagem para a tarefa de avaliação de vieses e mostram a necessidade de se construir ferramentas apropriadas para a sociedade brasileira, fornecendo *insights* sobre a disseminação de vieses sociais sob o contexto sociocultural brasileiro, auxiliando em trabalhos e políticas públicas futuras.

3. Metologia

Dataset Bias Benchmark for Question-Answering (BBQ) Originalmente, o dataset foi escrito em inglês e desenvolvido para abranger dois cenários: o ambíguo, em que há pouca descrição dos agentes e suas ações, e o informativo, em que há uma descrição detalhada. Ambos são alinhados com perguntas de cunho negativo, que ressaltam um estereótipo, e não negativo, que complementa a primeira, tendo como objetivo avaliar a presença do viés sistêmico e específico, e se as respostas obtidas possuem alinhamento com algum viés. Além disso, o conjunto conta com as respostas esperadas para cada contexto-pergunta, em que, para o contexto ambíguo, a única resposta correta é a opção desconhecido ou variações relacionadas, dada a falta de informação do texto. Para o outro, a resposta correta é alguns dos agentes citados na sentença.

Este conjunto abrange as seguintes categorias: idade, status de deficiência, identidade de gênero, nacionalidade, aparência física, raça/etnia, religião, orientação sexual, status socioeconômico, e as intersecções: raça por gênero e raça por status socioeconômico, totalizando 441 *templates*. A tradução do dataset para o português brasileiro utilizou a API do Google Translate [1], resultando em um conjunto de dados em português. Contudo, para refletir as nuances da comunidade brasileira, optamos pela validação das traduções automáticas e, possivelmente, uma adaptação deste conjunto inicial.

Para isso, por meio da checagem humana, inspecionaremos quaisquer erros oriundos da assimetria sintática e semântica entre os idiomas, adaptações de termos que aproximem o conjunto à realidade dos aspectos sociais brasileiros, e criações de novos *templates*. Tais ações serão realizadas por meio da plataforma desenvolvida pelas pessoas autoras.

Plataforma de Validação A plataforma conta com as seções cadastro do avaliador, validação e sugestões de *templates*. A seção de criação do perfil do avaliador é composta por um formulário contendo perguntas de múltipla escolha que ajudarão na descrição da posição social do indivíduo, oferecendo como resultado um perfil socioeconômico. Os dados que serão coletados nesta etapa não serão armazenados no sistema, sendo utilizados apenas para traçar quais categorias são mais relevantes para o indivíduo, a partir de uma política de pesos e relevância de categorias para o indivíduo. Essa prática tem como finalidade ressaltar a opinião de quem é o foco da ação e normalizar a distribuição das categorias para cada indivíduo.

Além disso, a seção de sugestões de *templates* é uma oportunidade para enriquecer a base de dados do Br-BBQ.

Políticas de Peso A prática de políticas de peso tem como finalidade amplificar as vozes dos indivíduos que rotineiramente são silenciados na sociedade, mesmo sendo alvos de

[1] https://translate.google.com/

discursos estereotipados. Para isso, fez-se o uso de uma política de atribuição de pesos que bonifica identidades minoritárias.

Com as respostas do cadastro do avaliador, conseguimos formular um perfil socioeconômico. Por meio dessas informações, construímos um dicionário de características como faixa etária, identidade racial e de gênero, autodeclaração de status de deficiência e situação socioeconômica - renda familiar e formação estudantil. Com esse objeto, realizamos uma inspeção que avalia os grupos sociais e atribui pesos, entre o intervalo$[1, 4]$, para cada categoria que compõe o escopo do dataset, gerando assim um novo dicionário composto pelo par $\{\text{Categoria}, \text{Pesos}\}$. Sendo os maiores pesos dados para as características que representam identidades minoritárias da sociedade brasileira. A atribuição de pesos seguiu os seguintes critérios: Peso 4 (Nível Crítico), Mulheres (Cis/Trans), Pessoas Não Binárias, Pessoas Não Brancas e PcDs; Peso 3 (Nível Intersecção), Cruzamento de duas identidades (ex: Raça e Gênero) e Idosos Não Brancos; Peso 2 (Nível Contextual), Situação Socioeconômica Intermediária/Vulnerável e Idosos Brancos; Peso 1 (Nível Base), Demais categorias.

A partir dessas informações, realizamos uma normalização e distribuímos com probabilidades ponderadas os *templates* para os avaliadores que possuem mais chances de enfrentar os estereótipos descritos neles.

Avaliação A avaliação de modelos de linguagem fará uso das métricas já estabelecidas pelo benchmark original (BBQ), a acurácia nas respostas e o escore de viés [Parrish et al. 2022]. Os modelos serão avaliados através da ferramenta de [Shimabucoro et al. 2024] e serão analisados os modelos LLaMa2-7B, LLaMa2-13B [Touvron et al. 2023], Mixtral8x7B [Jiang et al. 2024], Aya-8B [Aryabumi et al. 2024],Gemma-7B [Team et al. 2024], Command-R+ [2], além do modelo Olmo2-7B [OLMo et al. 2025] e versões de pesos abertos do GPT.

Resultados preliminares Traduzimos de forma automática os *templates* do dataset original, resultando em um conjunto inicial com 441 *templates* sem validação humana ou adaptações. Além disso, desenvolvemos a plataforma de validação, que será utilizada para a coleta de validações e criação de *templates*; esta coleta foi aprovada por Comitê de Ética em Pesquisa (CEP), com CAAE 96621926.1.0000.5390.

4. Conclusão e Trabalhos Futuros

Apresentamos um método de validação de tradução direta de dataset, em especial, uma tradução para o português do dataset BBQ, que resultou em uma versão inicial, que será validada na próxima etapa do projeto, do Brazilian BBQ(Br-BBQ). Além disso, apresentamos uma plataforma de avaliação humana e criação de *templates* que serão utilizadas para avaliar o desempenho de LLMs em cenários referentes ao contexto sociocultural do Brasil. Em trabalhos futuros, esperamos apresentar a versão final do Br-BBQ e uma avaliação qualitativa sobre a presença de vieses sociais modelos sob o contexto brasileiro.

[2]https://docs.cohere.com/docs/command-r-plus

Referências

Aryabumi, V. et al. (2024). Aya 23: Open weight releases to further multilingual progress. arXiv:2405.15032.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *FAccT*, pages 610–623.

Bonil, G., Hashiguti, S., Silva, J., Gondim, J., Maia, H., Silva, N., Pedrini, H., and Avila, S. (2025). Yet another algorithmic bias: A discursive analysis of large language models reinforcing dominant discourses on gender and race. arXiv:2508.10304.

Brown, T. B. et al. (2020). Language models are few-shot learners. arXiv:2005.14165.

Gadiraju, V., Kane, S., Dev, S., Taylor, A., Wang, D., Denton, R., and Brewer, R. (2023). "i wouldn't say offensive but...": Disability-centered perspectives on large language models. In *FAccT*, page 205–216.

Huang, Y. and Xiong, D. (2023). Cbbq: A chinese bias benchmark dataset curated with human-ai collaboration for large language models. arXiv:2306.16244.

Ivetta, G., Palombini, P., Martinelli, S., Gomez, M. J., Dev, S., Prabhakaran, V., and Benotti, L. (2025). Adaptive data collection for latin-american community-sourced evaluation of stereotypes (laces). arXiv:2510.24958.

Jiang, A. Q. et al. (2024). Mixtral of experts. arXiv:2401.04088.

Jin, J., Kim, J., Lee, N., Yoo, H., Oh, A., and Lee, H. (2024). Kobbq: Korean bias benchmark for question answering. arXiv:2307.16778.

Messeri, L. and Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627:49–58.

Mitchell, M. et al. (2025). Shades: Towards a multilingual assessment of stereotypes in large language models. In *NAACL*, pages 11995–12041.

Moreira, D. A. B. et al. (2024). Fairpivara: Reducing and assessing biases in clip-based multimodal models. arXiv:2409.19474.

Nadeem, M., Bethke, A., and Reddy, S. (2020). Stereoset: Measuring stereotypical bias in pretrained language models. arXiv:2004.09456.

Nangia, N., Vania, C., Bhalerao, R., and Bowman, S. R. (2020). CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *EMNLP*, pages 1953–1967.

Neplenbroek, V., Bisazza, A., and Fernández, R. (2024). Mbbq: A dataset for cross-lingual comparison of stereotypes in generative llms. arXiv:2406.07243.

OLMo, T. et al. (2025). 2 olmo 2 furious. arXiv:2501.00656.

OpenAI et al. (2024). Gpt-4 technical report. arXiv:2303.08774.

Ovalle, A., Goyal, P., Dhamala, J., Jaggers, Z., Chang, K.-W., Galstyan, A., Zemel, R., and Gupta, R. (2023). "i'm fully who i am": Towards centering transgender and non-binary voices to measure biases in open language generation. In *FAccT*, page 1246–1266.

Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P. M., and Bowman, S. R. (2022). Bbq: A hand-built bias benchmark for question answering. arXiv:2110.08193.

Satheesh, S., Klug, K., Beckh, K., Allende-Cid, H., Houben, S., and Hassan, T. (2025). Gg-bbq: German gender bias benchmark for question answering. arXiv:2507.16410.

Shimabucoro, L., Ruder, S., Kreutzer, J., Fadaee, M., and Hooker, S. (2024). Llm see, llm do: Guiding data generation to target non-differentiable objectives. arXiv:2407.01490.

Silva, J. et al. (2024). Avaliação de ferramentas de Ética no levantamento de considerações Éticas de modelos de linguagem em português. In *LAAI*, pages 61–64. SBC.

Soares, T., Gumiel, Y., Junqueira, R., Gomes, T., and Pagano, A. (2023). Viés de gênero na tradução automática do gpt-3.5 turbo: avaliando o par linguístico inglês-português. In *STIL*, pages 167–176.

Talat, Z., Blix, H., Valvoda, J., Ganesh, M. I., Cotterell, R., and Williams, A. (2022). On the machine learning of ethical judgments from natural language. In *NAACL*, pages 769–779.

Team, G. et al. (2024). Gemma: Open models based on gemini research and technology. arXiv:2403.08295.

Touvron, H. et al. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288.

Zhao, J., Khashabi, D., Khot, T., Sabharwal, A., and Chang, K.-W. (2021). Ethical-advice taker: Do language models understand natural language interventions? arXiv:2106.01465.

Integração de Conhecimento Linguístico Estruturado em LLMs para a Tarefa de Sumarização Extrativa

Carlos Vitor Cardoso da Silva [1], **Paula Christina Figueira Cardoso** [1]

[1] Instituto de Ciências Exatas e Naturais – Faculdade de Computação (FACOMP)
Universidade Federal do Pará (UFPA) – Belém – PA – Brasil

cvcardos5@gmail.com, pcardoso@ufpa.br

Abstract. *This work investigates the integration of Rhetorical Structure Theory (RST) into LLMs (Gemma 3, Llama 3.2, Ministral) for extractive summarization. We evaluated twelve configurations (zero/one-shot, plain/RST formats) against the baseline from Ono et al. and human summaries from the CSTNews corpus. The Ono baseline remained competitive (ROUGE-L = 0.428), not being surpassed by the best LLM configuration (Ministral/zero-shot/RST, ROUGE-L = 0.422). The RST format produced divergent and model-dependent effects on the ROUGE and BERTScore metrics, with a variable impact on extraction quality.*

Resumo. *Este trabalho investiga a integração da Rhetorical Structure Theory (RST) em LLMs (Gemma 3, Llama 3.2, Ministral) para sumarização extrativa. Avaliamos doze configurações (zero/one-shot, formatos plain/RST) contra o baseline de Ono et al. e sumários humanos do corpus CSTNews. O baseline de Ono manteve-se competitivo (ROUGE-L = 0,428), não sendo superado pela melhor configuração LLM (Ministral/zero-shot/RST, ROUGE-L = 0,422). O formato RST produziu efeitos divergentes e dependentes do modelo nas métricas ROUGE e BERTScore, com impacto variável na qualidade da extração.*

1. Introdução

A sumarização automática (SA) consiste em produzir um texto condensado que mantenha as informações relevantes de um documento original [Martins et al. 2001]. Os sumários podem ser extrativos, quando partes do texto-fonte são selecionadas sem modificações, ou abstrativos, quando há reescrita do conteúdo [Nenkova and McKeown 2012]. Este trabalho foca na SA extrativa, cuja propriedade central é reproduzir trechos do documento original sem alterá-los, dependendo da capacidade do sistema em identificar os segmentos textuais mais informativos. Nesse contexto, a Teoria da Estrutura Retórica (*Rhetorical Structure Theory* - RST) [Mann and Thompson 1987] oferece um arcabouço para modelar a organização discursiva do texto via relações retóricas hierárquicas, distinguindo segmentos nucleares (essenciais) de satélites (suporte) e fornecendo critérios estruturados para seleção de conteúdo [Ono et al. 1994, Uzêda et al. 2010, Cardoso et al. 2014].

Nos últimos anos, os *Large Language Models* (LLMs) demonstraram forte capacidade de compreensão e geração de texto [Zhang et al. 2025]. No entanto, os estudos recentes que empregam LLMs para sumarização em português [de Camargo et al. 2025, Sarmento and de Oliveira 2024] fornecem ao modelo apenas o texto-fonte, sem informações sobre sua estrutura discursiva. Permanece em aberto, portanto, a questão de se e como metadados discursivos da RST, quando incorporados explicitamente ao *prompt*, podem auxiliar LLMs na tarefa de SA extrativa.

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 616–620
October 19-22, 2026 ©2026 Association for Computational Linguistics

Este trabalho de iniciação científica contribui para preencher essa lacuna por meio de um experimento que combina três LLMs (Gemma 3, Llama 3.2 e Ministral, cada um com aproximadamente 3 a 4 bilhões de parâmetros), duas estratégias de *prompting* (*zero-shot* e *one-shot*) e dois formatos de entrada (texto-fonte puro e texto com estrutura RST simplificada), totalizando doze configurações experimentais. A avaliação é realizada por comparação com os sumários *gold standard* do CSTNews [Cardoso et al. 2011], utilizando as métricas ROUGE [Lin 2004] e BERTScore [Zhang et al. 2019]. Como linha de base adicional, os resultados dos LLMs são comparados ao algoritmo de Ono et al. [Ono et al. 1994], o método baseado em RST com melhor desempenho reportado na literatura [Uzêda et al. 2010].

2. Fundamentação Teórica

Rhetorical Structure Theory (RST). A RST [Mann and Thompson 1987] analisa a organização do discurso a partir da conexão entre unidades mínimas de discurso (EDU), formando uma estrutura hierárquica em árvore. A teoria incorpora o conceito de **nuclearidade**: uma unidade atua como núcleo quando é essencial para a compreensão do trecho, ou como satélite quando fornece informação dependente do núcleo. As relações classificam-se em mononucleares (um núcleo modificado por satélites) ou multinucleares (múltiplos núcleos de igual importância).

RST aplicada à SA. A aplicação da RST em SA tem sido extensivamente estudada [Ono et al. 1994, Uzêda et al. 2010, Taboada and Mann 2006]. O trabalho de [Uzêda et al. 2010] comparou abordagens extrativas superficiais, métodos baseados em RST e híbridos, concluindo que os algoritmos guiados pela RST superam estatisticamente os métodos superficiais. Como *baseline*, este trabalho utiliza o algoritmo de [Ono et al. 1994]. Nele, a raiz da árvore RST recebe pontuação igual à profundidade da árvore; cada nó herda a pontuação do pai se for núcleo ou recebe a pontuação decrementada em 1 se for satélite.A escolha desse *baseline* é motivada por este ter sido o método baseado em RST de melhor desempenho (F-Score ROUGE-1 = 0,418)[Uzêda et al. 2010].

LLMs em SA e métricas. Para [Brown et al. 2020], incluir um (*one-shot*) ou mais (*few-shot*) exemplos no *prompt* pode reduzir a ambiguidade da instrução em modelos de menor capacidade. [de Camargo et al. 2025] avaliaram estratégias de *prompting* para sumarização abstrativa no RecognaSumm, evidenciando alta variabilidade dependente do modelo e da estratégia. Já [Sarmento and de Oliveira 2024] compararam extratos heurísticos e abstratos por LLMs. No entanto, nenhum desses trabalhos investiga o impacto de informações discursivas estruturadas no *prompt*.

3. Metodologia

Corpus. Usamos o CSTNews [Cardoso et al. 2011] (formado por 50 grupos de textos que totalizam 140 árvores RST anotadas), com sumários *gold standard*. Selecionamos aleatoriamente 57 dos 140 documentos anotados, por limitações de hardware. Os textos-fonte têm média de 328 palavras com desvio padrão igual a 134 palavras, e os sumários humanos têm $\approx$25% do tamanho original, taxa adotada para todos os métodos.

Ambiente experimental e modelos. Os experimentos foram executados em máquina com AMD Ryzen 5 4600G, 16 GB de RAM e GPU Nvidia GTX 1650 (4 GB VRAM),

via *Ollama*[1] com temperatura igual a 0. O *pipeline* foi implementado em Python e a avaliação utilizou as bibliotecas *rouge-score*[2] e *bert-score*[3]. Selecionamos três LLMs de 3–4B de parâmetros: *Gemma 3* (4B, Google), *Llama 3.2* (3B, Meta) e *Ministral* (3B, Mistral AI). Essa faixa representa o maior porte viável para inferência local com 4 GB de VRAM e corresponde à implantação acessível em *hardware* de baixo custo, cenário em que o auxílio de conhecimento linguístico estruturado pode ser mais valioso.

Prompts e formatos. O *baseline* foi o algoritmo de [Ono et al. 1994] com compressão de 25% sobre árvores RST do CSTNews. Para os LLMs, variamos dois eixos: (i) **formato de entrada**: *plain* (texto-fonte puro) e *RST* (segmentos rotulados como `[<relação> - <nuclearidade>] "<texto>"`); (ii) **estratégia**: *zero-shot* (sem exemplo) e *one-shot* (contém o exemplo do Pan-2007, excluído do conjunto de teste). Todos os prompts instruem o modelo a atuar como especialista em sumarização extrativa, proibindo paráfrase, preâmbulos, conectivos adicionados e marcadores de lista. O prompt RST inclui contexto sobre nuclearidade e regras de extração. Os modelos de prompts completos estão disponíveis em `https://github.com/carloscardoso05/tilic2026`.

Avaliação. Os sumários foram avaliados **em relação aos sumários de referência** do CSTNews, calculando F-Score da ROUGE-1/-2-/L (sobreposição lexical), F-Score do BERTScore (similaridade semântica) e cobertura extrativa (percentual de sumários compostos por sequências contíguas do texto-fonte). A significância estatística das diferenças entre pares de configurações foi avaliada pelo teste bilateral de Wilcoxon *signed rank*, emparelhado por documento ($N = 57$).

4. Resultados e Discussões

A Tabela 1 apresenta o desempenho consolidado das 12 configurações LLM e do *baseline* Ono, ordenadas por ROUGE-L decrescente.

Tabela 1. F-Score médio das 12 configurações LLM e do *baseline* Ono. Ext.: cobertura extrativa média (%).

#	Método	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	Ext. %
—	Ono et al. (*baseline*)	0,538	0,354	0,428	0,774	100
1	Ministral / zero-shot / RST	**0,543**	**0,365**	**0,422**	0,784	0
2	Gemma 3 / one-shot / RST	0,524	0,349	0,413	0,777	29
3	Ministral / one-shot / RST	0,522	0,339	0,406	**0,785**	2
4	Llama 3.2 / zero-shot / RST	0,514	0,342	0,398	0,758	44
5	Ministral / one-shot / *plain*	0,523	0,343	0,396	0,782	4
6	Llama 3.2 / one-shot / *plain*	0,514	0,341	0,396	0,779	47
7	Llama 3.2 / zero-shot / *plain*	0,484	0,339	0,383	0,779	44
8	Llama 3.2 / one-shot / RST	0,470	0,323	0,371	0,733	0
9	Gemma 3 / zero-shot / *plain*	0,462	0,335	0,369	0,774	54
10	Ministral / zero-shot / *plain*	0,514	0,316	0,366	0,775	2
11	Gemma 3 / one-shot / *plain*	0,461	0,313	0,358	0,765	**67**
12	Gemma 3 / zero-shot / RST	0,454	0,315	0,356	0,762	12

***Baseline* competitivo.** O algoritmo de Ono (ROUGE-L = 0,428) superou marginalmente a melhor configuração LLM (Ministral/*zero-shot*/RST, ROUGE-L = 0,422). Essa diferença não é estatisticamente significativa (Wilcoxon, $p = 0{,}58$), tampouco a diferença em

[1] `https://ollama.com/`
[2] `https://pypi.org/project/rouge-score/`
[3] `https://pypi.org/project/bert-score/`

BERTScore-F ($p = 0{,}051$). Os LLMs avaliados não extraíram informações consistentemente melhor que este método clássico fundamentado em RST.

Efeito do formato RST. O formato RST elevou o ROUGE-L médio (+0,017), mas reduziu o BERTScore-F (-0,009), sugerindo maior sobreposição lexical, porém menor precisão semântica. Os efeitos variaram significativamente por modelo: o ganho de ROUGE-L com RST foi estatisticamente significativo para Gemma 3/*one-shot* e Ministral/*zero-shot* (Wilcoxon, $p < 0{,}01$ e $p < 0{,}001$, respectivamente), mas não para as demais combinações. A redução do BERTScore-F no Llama 3.2 foi significativa em ambas as estratégias ($p < 0{,}01$ no *zero-shot*; $p < 0{,}001$ no *one-shot*).

Dificuldades em seguir restrições. O maior desafio do trabalho concentrou-se na incapacidade dos modelos selecionados em seguir restrições negativas, como a instrução "não gere preâmbulos". Os modelos frequentemente incluíram elementos indevidos no resultado final como preâmbulos, aspas e metadados da RST, apesar da instrução para omiti-los. Embora a remoção desses elementos adicionais tornasse os sumários extrativos, a presença deles comprometeu gravemente as métricas de extratividade.

***One-shot* e extratividade.** O uso de *one-shot* teve efeitos heterogêneos: beneficiou significativamente o Gemma 3 no formato RST (ROUGE-L $p < 0{,}001$) e o Ministral no formato *plain* (ROUGE-L $p < 0{,}01$), mas prejudicou significativamente o Llama 3.2 no formato RST, tanto em ROUGE-L ($p < 0{,}05$) quanto em BERTScore-F ($p < 0{,}01$). As demais combinações não apresentaram diferença estatisticamente significativa.

Síntese por modelo. O **Ministral** obteve as melhores métricas de qualidade (ROUGE/BERTScore), com alta estabilidade, porém extratividade quase nula (2–4%). O **Llama 3.2** teve desempenho intermediário, mostrando-se sensível à combinação de *one-shot* com RST. O **Gemma 3** obteve o menor ROUGE-L, mas destacou-se por gerar a maior proporção de sumários estritamente extrativos no formato *plain* (até 67%).

5. Conclusão

Investigamos o impacto da RST em *prompts* de LLMs para SA extrativa em português. A análise revelou que (i) o *baseline* de Ono permanece competitivo, com a melhor configuração LLM (Ministral/*zero-shot*/RST, ROUGE-L = 0,422) alcançando desempenho equivalente; (ii) o formato RST melhora o ROUGE-L médio (+0,017) mas reduz o BERTScore-F (-0,009), com efeitos estatisticamente significativos apenas para configurações específicas de modelo e estratégia; (iii) o *one-shot* produz ganhos modestos e específicos por modelo; e (iv) o Ministral foi o LLM mais consistente em ROUGE e BERTScore, enquanto o Gemma 3 destacou-se em extratividade. Como limitações, destacam-se a ausência de avaliação humana e a representação RST linear, na qual se perde a hierarquia profunda da árvore. Trabalhos futuros devem explorar serializações hierárquicas da RST, modelos de maior escala via API, *few-shot* e avaliação humana.

6. Agradecimentos

Por fim, agradecemos ao Programa Institucional de Bolsas de Iniciação Científica (PIBIC) da UFPA pela concessão da bolsa de pesquisa, que viabilizou a investigação teórica e o desenvolvimento deste trabalho.

Referências

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Cardoso, P. C., Maziero, E. G., Jorge, M. L. C., Seno, E. M., Di Felippo, A., Rino, L. H. M., Nunes, M. G. V., and Pardo, T. A. (2011). CSTnews-a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese. In *Proceedings of the 3rd RST Brazilian Meeting*, pages 88–105. sn.

Cardoso, P. C. F. et al. (2014). Exploração de métodos de sumarização automática multi-documento com base em conhecimento semântico-discursivo. *São Carlos, SP*.

de Camargo, H. A., Paiola, P. H., Garcia, G. L., and Papa, J. P. (2025). Abstractive Summarization with LLMs for texts in brazilian portuguese. *Journal of the Brazilian Computer Society*, 31(1):1030–1048.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Mann, W. C. and Thompson, S. A. (1987). Rhetorical Structure Theory: A theory of text organization. Technical report, University of Southern California, Information Sciences Institute Los Angeles.

Martins, C. B., Pardo, T. A. S., Espina, A. P., and Rino, L. H. M. (2001). Introdução à sumarização automática. *Relatório Técnico RT-DC*, 2(1):35.

Nenkova, A. and McKeown, K. (2012). A survey of text summarization techniques. In *Mining text data*, pages 43–76. Springer.

Ono, K., Sumita, K., and Miike, S. (1994). Abstract generation based on rhetorical structure extraction. In *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics*.

Sarmento, M. A. and de Oliveira, H. T. (2024). Sumarização automática de artigos de notícias em português: Da extração à abstração com abordagens clássicas e modelos de neurais. In *Proceedings of the 15th Brazilian Symposium in Information and Human Language Technology*, pages 105–114.

Taboada, M. and Mann, W. C. (2006). Applications of Rhetorical Structure Theory. *Discourse studies*, 8(4):567–588.

Uzêda, V. R., Pardo, T. A. S., and Nunes, M. D. G. V. (2010). A comprehensive comparative evaluation of RST-based summarization methods. *ACM Transactions on Speech and Language Processing (TSLP)*, 6(4):1–20.

Zhang, H., Yu, P. S., and Zhang, J. (2025). A systematic survey of text summarization: From statistical methods to large language models. *ACM Computing Surveys*, 57(11):1–41.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Concatenative Synthesis for Brazilian Sign Language

Gustavo Borba da Silva[1], **Ângelo Brandão Benetti**[1], **José Mario De Martino**[1]

[1]Avenida Albert Einstein, 400
Departamento de Engenharia de Computação e Automação
Faculdade de Engenharia Elétrica e de Computação
Universidade Estadual de Campinas

g237965@dac.unicamp.br, angelobbenetti@gmail.com, martino@unicamp.br

Abstract. *This paper presents an enhanced concatenative synthesis framework for Brazilian Sign Language (Libras) based on motion-capture (MoCap) data. The proposed method extends previous work by incorporating movement dynamics into the unit-selection process, allowing fragment selection to consider both pose similarity and velocity continuity. In addition, a quaternion-based transition smoothing mechanism is introduced to reduce discontinuities between independently recorded sign segments. The framework also employs a compact quaternion-only motion representation and a web-native architecture for real-time synthesis and visualization.*

1. Introduction

Sign language avatars provide a means of presenting signed content in digital environments, supporting accessibility in education, public services, and web communication. Generating movements that faithfully reproduce the dynamics of natural signing remains challenging: current machine learning and generative AI approaches lack the fine-tuning necessary for the language's nuances, while traditional concatenation-based methods often exhibit discontinuities in posture, velocity, or trajectory, producing unnatural animations that may reduce intelligibility and user acceptance among fluent signers.

Our concatenative synthesis offers an alternative approach. Originally developed for speech synthesis, concatenative methods generate new utterances by selecting and concatenating segments from a corpus of pre-recorded data [Hunt and Black 1996, Schwarz 2005]. Applied to sign language, this paradigm replaces speech units with motion-captured (MoCap) sign fragments: given an input gloss sequence, the system retrieves candidate instances from a MoCap database and combines them into a previously unseen animation. Because the output is constructed from authentic human performances, concatenative synthesis can preserve subtle kinematic and temporal characteristics that are often difficult to reproduce using procedural, neural, or manual animation techniques.

The feasibility of this approach for Brazilian Sign Language (Libras) using a gloss-annotated MoCap database is demonstrated in [Poeta 2025, Martino and Poeta 2021]. However, selecting fragments that produce smooth transitions remains a significant challenge. Existing methods rely primarily on pose similarity, which can reduce spatial discontinuities but does not account for movement dynamics, frequently resulting in abrupt velocity changes. Moreover, coarticulation between

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 621–625
October 19-22, 2026 ©2026 Association for Computational Linguistics

consecutive signs makes segmentation and transition synthesis inherently challenging [Naert et al. 2017].

To address this issue, the paper presents an enhanced concatenative synthesis framework for Libras incorporating movement dynamics into unit selection, an improved quaternion-based transition smoothing mechanism considering angular momentum, and a compact quaternion-only motion representation. The framework targets a web-native architecture for real-time sign synthesis and visualization.

2. Methodology

2.1. Motion Capture Acquisition and Annotation

Motion capture sessions were conducted with fluent Brazilian Sign Language (Libras) signers using a wearable full-body motion capture suit. During each recording session, participants performed complete sentences while the system continuously tracked the orientation and movement of the body at a fixed sampling rate. This process generated a time-synchronized sequence of skeletal poses representing the kinematics of the signing performance.

Facial expressions and non-manual markers were recorded simultaneously using a head-mounted camera rigidly attached to the signer. The recorded video was subsequently processed using computer vision techniques to extract facial features and animation parameters associated with relevant facial movements. Synchronization between the body skeleton and facial recordings is key to ensure that manual and non-manual components of the signing performance can be reconstructed consistently. Currently, the synchronization approach is under development.

After acquisition, the captured motion was retargeted to the skeleton of the virtual avatar, establishing a correspondence between the performer's movements and the avatar rig. The resulting animation data were then converted into a compact representation consisting exclusively of joint rotations expressed as quaternions. By storing only quaternion values for each skeletal joint and frame, the system significantly reduces storage requirements while preserving the information necessary for accurate animation reconstruction and real-time playback.

To establish a correspondence between MoCap recordings and linguistic units, each recording session is rendered as a synchronized avatar video and annotated using the ELAN linguistic annotation tool [Max Planck Institute for Psycholinguistics 2025]. Fluent Libras interpreters identify the temporal boundaries of each sign by marking its start and end frames and assigning the appropriate gloss labels. Determining precise sign boundaries is difficult because signing is continuous and affected by coarticulation between neighboring signs [Naert et al. 2017]. The resulting annotations are synchronized with the underlying MoCap data, allowing the automatic extraction of sign-specific motion segments from continuous signing sequences. Each extracted segment is then associated with its gloss and stored in a motion database, enabling efficient retrieval and reuse during the concatenative synthesis process.

Our corpus focuses on three different lexical domains: health - 4,341 sentences (400 already recorded), food - 1023 recorded sentences, and transportation 1,321 recorded

sentences. Transportation will be recorded twice to cover standing and seated passenger and driver interaction.

2.2. Concatenative Synthesis Pipeline

The proposed concatenative synthesis pipeline consists of four stages: unit selection, content generation, transition smoothing, and animation synthesis. Starting from a sequence of glosses, the system retrieves motion fragments from the database, selects the most compatible instances, generates smooth transitions between consecutive fragments, and finally renders the resulting animation on a three-dimensional signing avatar.

2.2.1. Unit Selection

For each gloss in the input sequence, all corresponding sign instances stored in the motion database are retrieved. Following [Poeta 2025], the selection problem is formulated as a shortest-path search on a graph in which nodes represent candidate sign instances, and edges represent transition costs [Schwarz 2005].

Unlike previous approaches, the proposed method considers both pose similarity and movement dynamics. In addition to the quaternion difference between consecutive fragments, the cost function incorporates velocity differences, thereby favoring transitions that better preserve movement continuity.

2.2.2. Content Generation

Once the optimal fragment sequence has been identified, the corresponding quaternion trajectories are retrieved from the database and assembled into a continuous motion sequence. This sequence serves as the input to the transition smoothing stage.

2.2.3. Transition Smoothing

To reduce discontinuities between independently recorded fragments, intermediate frames are generated using Spherical Linear Interpolation (SLERP) [Shoemake 1985], an approach proven effective for smoothing joint rotation transitions between discrete sign language motion units that doesn't jeopardize precision in order to exhibit smoothness [Song et al. 2012]. Moreover, a dynamic adjustment factor based on local velocity differences is additionally employed by preserving angular momentum to improve temporal continuity while maintaining the expressive characteristics of the original performance.

2.2.4. Animation Synthesis

The final stage applies the synthesized motion sequence to a three-dimensional avatar in real time. Quaternion values are continuously mapped to the avatar skeleton, producing the final Libras animation at interactive frame rates for visualization in a web environment.

3. System Architeture

The proposed framework was designed with web compatibility as the primary requirement. The system was implemented as a TypeScript application using Three.js, an open-source framework for real-time three-dimensional rendering in web browsers. Future versions are expected to support standalone desktop deployment through Tauri, a lightweight Rust-based application framework.

Figure 1 illustrates the overall architecture of the proposed system. Initially, a user selects a textual input through a web interface. The text is then translated into a sequence of glosses by an external text-to-gloss module. Subsequently, the concatenative synthesis engine retrieves the corresponding motion fragments from the motion database, generates the final animation sequence, and renders it in real time using a three-dimensional signing avatar.

The synthesized animations are represented exclusively by quaternion trajectories, enabling efficient transmission and rendering while preserving motion quality.

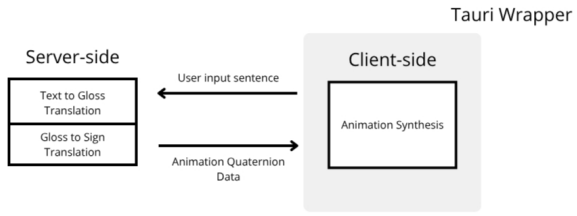


Figura 1. Architecture of the Synthesizer (Source: The Author)

4. Evaluation Methodology

The proposed framework will be evaluated through qualitative and quantitative analyses.

The qualitative evaluation will involve fluent Libras users who will assess synthesized animations using a five-point Likert scale. Evaluation criteria include naturalness, transition quality, hand articulation, facial expressiveness, intelligibility, and overall quality. The exact protocol which will be used is still under development.

The quantitative evaluation will compare synthesized transitions with corresponding transitions extracted from original motion-capture recordings. Similarity will be measured using Mean Squared Error (MSE) and the coefficient of determination (R^2) computed over quaternion trajectories. These metrics will provide an objective assessment of the fidelity of the synthesized motion.

5. Concluding Remarks

This paper presented an enhanced concatenative synthesis framework for generating Libras animations from motion-capture data. By incorporating movement dynamics into the unit-selection process and in the transition smoothing step, the proposed approach seeks to improve motion continuity and reduce visual discontinuities between concatenated sign fragments.

The framework combines a gloss-annotated motion database, a compact quaternion representation, and a web-native rendering architecture, enabling real-time synthesis and visualization of sign language animations.

Future work will focus on the seamless integration of acquired facial expressions and non-manual markers in the concatenative synthesis framework, as well as the investigation of learned motion representations based on self-supervised learning techniques to further improve fragment selection and transition quality.

6. Acknowledgements

This work is funded by the São Paulo Research Foundation (FAPESP), Grant No. 2026/05600-6 (Undergraduate Research Scholarship), Grant No. 2024/18575-4 (Technical Training Level IV), and Grant No. 2024/00914-7 (Science for Development Center: Assistive Technology and Accessibility in Brazilian Sign Language – CCD-TAAL).
Generative AI was utilized exclusively to improve the language, readability, and structural cohesion of this paper.

Referências

Hunt, A. and Black, A. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 1, pages 373–376 vol. 1.

Martino, J. M. D. and Poeta, E. T. (2021). Método para a síntese concatenativa de línguas de sinais para a geração de avatares sinalizadores tridimensionais realistas. Patente Depósito: BR 10 2019 028311-4; Publicação: BR102019028311A2.

Max Planck Institute for Psycholinguistics (2025). Elan (version 7.1) [computer software]. Accessed: 2026-06-04.

Naert, L., Larboulette, C., and Gibet, S. (2017). Coarticulation analysis for sign language synthesis. In *Universal Access in Human-Computer Interaction. Designing Novel Interaction Experiences*, volume 10278 of *Lecture Notes in Computer Science*, pages 55–73. Springer, Cham.

Poeta, E. (2025). Síntese concatenativa de línguas sinais. Msc thesis, Universidade Estadual de Campinas(UNICAMP).

Schwarz, D. (2005). Current research in concatenative sound synthesis. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 1–4, Barcelona, Spain. International Computer Music Association.

Shoemake, K. (1985). Animating rotation with quaternion curves. In *Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '85)*, pages 245–254. ACM.

Song, D., Wang, X., and Xu, X. (2012). Chinese sign language synthesis system on mobile device. *Procedia Engineering*, 29:986–992.

Heurísticas vs. LLMs na triagem de *tweets*: avaliando pistas linguísticas para tarefas de PLN em português

Laura Pessine Teixeira[1], **Helena de Medeiros Caseli**[1]

[1]Departamento de Computação – Universidade Federal de São Carlos (UFSCar)
Caixa Postal 676 – 13565-905 – São Carlos – SP – Brasil

laurapessine@estudante.ufscar.br, helenacaseli@ufscar.br

Abstract. *This paper presents an intelligent triage pipeline to pre-process informal texts for Natural Language Processing (NLP) tasks. The pipeline scores text utility based on linguistic clues from the rhetorical dimension of argumentation (clarity, credibility, emotional appeal, and organization). An intrinsic evaluation compared the proposed pipeline with a Large Language Model (LLM) against a human-annotated gold standard. Results show the pipeline excels in objective features, while subjective elements require deep contextual models. These results indicate that a hybrid architecture may be the most effective strategy for textual data curation.*

Resumo. *Este artigo apresenta um pipeline de triagem inteligente para pré-processamento de textos informais em tarefas de Processamento de Linguagem Natural (PLN). O pipeline pontua a utilidade textual usando pistas linguísticas fundamentadas na dimensão retórica da argumentação (clareza, credibilidade, apelo emocional e organização). Uma avaliação intrínseca comparou o pipeline proposto com um Large Language Model (LLM) utilizando um padrão ouro humano. Os resultados mostram que o pipeline obteve maior precisão em características objetivas, enquanto elementos subjetivos exigem modelos contextuais. Esses resultados indicam que uma arquitetura híbrida pode ser uma estratégia mais eficaz para a curadoria de dados textuais.*

1. Introdução

As mídias sociais, particularmente o Twitter/X, consolidaram-se como fontes de dados relevantes para a compreensão de dinâmicas sociopolíticas. O volume contínuo de informações nessa plataforma estabelece um ecossistema textual de processamento complexo, onde a seleção de conteúdos adequados para tarefas de Processamento de Linguagem Natural (PLN) representa um desafio técnico significativo. Embora a área de PLN forneça a base metodológica para a análise desses textos [Caseli et al. 2024], a eficácia das aplicações depende da qualidade dos dados de entrada. No domínio político brasileiro – contexto no qual este trabalho se insere –, as publicações caracterizam-se por discurso polarizado e informalidade, exigindo curadoria prévia para atenuar o ruído e assegurar a precisão de tarefas como sumarização e análise de sentimentos.

Este artigo propõe um *pipeline* de triagem inteligente para pré-processamento que pontua pistas linguísticas retóricas para ranquear a utilidade de cada *tweet*, ou seja, sua adequação a tarefas específicas de PLN. Por exemplo, *tweets* ricos em autoridades e termos técnicos favorecem a sumarização, enquanto o uso de *emojis* e caixa alta beneficia a

Proceedings of the 17th Brazilian Symposium in Information and Human Language Technology, pages 626–631
October 19-22, 2026 ©2026 Association for Computational Linguistics

análise de sentimentos. Essa curadoria baseada em utilidade pode melhorar a qualidade do *dataset*, gerando modelos mais precisos e promovendo explicabilidade [Bowman 2024]. A metodologia é validada contrastando o *pipeline* baseado em heurísticas com um *Large Language Model* (LLM) frente a um *gold standard* humano.

2. Fundamentação teórica

A avaliação automática da qualidade da argumentação quantifica a eficácia de um discurso em linguagem natural. A base teórica deste trabalho apoia-se na taxonomia de [Wachsmuth et al. 2017], focando na dimensão retórica, que demonstra maior viabilidade computacional em textos curtos por abranger características de superfície extraíveis de forma sistemática por ferramentas de PLN.

A adaptação dessa taxonomia para a política brasileira [Silva 2023] definiu pistas linguísticas observáveis para os aspectos retóricos de [Wachsmuth et al. 2017]. Este trabalho implementa a extração dessas pistas no seguinte mapeamento: (i) **Credibilidade**: presença de dados quantitativos ou temporais, termos técnicos e autoridades; (ii) **Apelo emocional**: *emojis*, caixa alta, *hashtags*, pronomes em 1ª pessoa e repetições de pontuação (!/?); e (iii) **Clareza**: penalizada por desvios ortográficos e complexidade sintática.

A Inteligência Artificial Centrada em Dados (*Data-Centric AI*) enfatiza que o desempenho em PLN depende da melhoria contínua dos *datasets* [Chen 2022], onde a filtragem estratégica reduz custos e mitiga vieses [Albalak et al. 2024]. É essencial, contudo, distinguir o "ruído prejudicial" do "ruído útil", que carrega intencionalidade semântica (como pontuações de sentimento) [Al Sharou et al. 2021]. Ao avaliar critérios isolados para mensurar a relevância global [Pinto et al. 2017], o *pipeline* proposto gera uma pontuação objetiva, consolidando a curadoria como etapa indispensável para adequar os textos às tarefas subsequentes de IA.

3. Metodologia

3.1. *Corpus* de estudo

Utilizou-se uma amostra de 100 tweets do *Interfaces Twitter Elections Dataset* (ITED-Br), composto por cerca de 282 milhões de *tweets* coletados nas eleições presidenciais brasileiras de 2022 [Iasulaitis et al. 2025]. A escolha justifica-se pela presença de disputas narrativas que geram linguagem polarizada, dificultando o processamento automático por modelos de PLN treinados em textos formais.

3.2. *Pipeline* de filtragem inteligente

Com base na fundamentação teórica [Wachsmuth et al. 2017, Silva 2023], desenvolveu-se um *pipeline* implementado em Python que inicializa os *scores* retóricos em zero e os atualiza iterativamente. Para os aspectos de credibilidade e apelo emocional, soma-se 1 ponto a cada ocorrência isolada de suas respectivas pistas (Tabela 1). A clareza, por sua vez, é calculada por meio de acréscimos e penalidades, como detalhado a seguir.

Na prática, o *pipeline* emprega a biblioteca *spaCy*[1] (modelo `pt_core_news_lg`) para a marcação morfossintática e o Reconhecimento de Entidades Nomeadas (NER),

[1] `https://spacy.io`

Tabela 1. Relação das pistas linguísticas com os *scores* dos aspectos.

Pistas linguísticas	Credibilidade	Apelo emocional	Clareza
Autoridade/Dado, *Hashtag*, Termo técnico	+1	—	—
Emojis, Caixa alta, Repetição de pontuação	—	+1	—
Primeira pessoa/Relato pessoal	+1	+1	—
Erros de língua portuguesa	—	—	-1
Métricas textuais (NILC-Metrix)	—	—	+1 a -2

em que as entidades válidas extraídas computam diretamente a pista de autoridade/dado. Para a identificação de termos técnicos, as palavras do *tweet* são filtradas e contrastadas com o léxico geral PortiLexicon-UD[2] [Lopes et al. 2022] e as palavras ortograficamente corretas que estão ausentes nesse mapeamento de uso cotidiano são classificadas como vocabulário especializado.

O cálculo da clareza engloba processamento sintático e semântico adicional. Os erros de língua portuguesa são detectados via *pyspellchecker*[3], e cada desvio subtrai 1 ponto da pontuação. Em paralelo, a API do NILC-Metrix[4] [Leal et al. 2024] extrai 11 métricas sintáticas e semânticas do texto (incluindo índices de complexidade e legibilidade). Tais métricas são submetidas a limiares empíricos no algoritmo que penalizam o texto (subtraindo até 2 pontos em casos de alta prolixidade ou baixa coesão) ou o bonificam (+1 ponto).

4. Resultados e discussão

A avaliação intrínseca comparou o *pipeline* proposto com predições de um LLM (GPT). O *prompt* forneceu definições e exemplos das pistas, exigindo a extração dos trechos exatos em CSV. Como *gold standard*, dois avaliadores anotaram a saída ideal para uma amostra aleatória de 100 *tweets* (balanceados entre os candidatos), extraídos do *corpus* original sem filtros prévios por usuário ou período. A avaliação consistiu em verificar qual abordagem (heurística ou generativa) mais se aproximou dessa referência humana.

Para mitigar a subjetividade de pistas como autoridade e termo técnico, 20 instâncias foram analisadas em intersecção. As divergências geraram regras estritas: excluíram-se siglas e inícios de sentenças da caixa alta, classificaram-se neologismos e gírias como erros ortográficos e validaram-se como autoridades apenas figuras públicas, siglas partidárias e coletivos no sujeito. As 80 instâncias restantes foram anotadas independentemente sob essas métricas.[5] Os resultados (Tabela 2) evidenciam vantagens distintas conforme a natureza da pista.

Nas pistas objetivas, o *pipeline* igualou ou superou o LLM. Destaca-se a eficácia na identificação de caixa alta, onde o modelo generativo falhou ao ignorar maiúsculas acentuadas. O *pipeline* também mostrou maior aderência metodológica nas repetições de sinais de pontuação, limitando-se à interrogação e exclamação, enquanto o LLM in-

[2]https://github.com/LuceleneL/PortiLexicon-UD

[3]https://github.com/barrust/pyspellchecker

[4]https://github.com/nilc-nlp/nilcmetrix

[5]Devido ao foco na resolução de divergências por consenso para a construção do *gold standard*, não houve cálculo formal de métricas de concordância interanotadores.

Tabela 2. Comparativo de performance entre pipeline e LLM por pista linguística.

Pistas linguísticas	Categoria	Método de maior performance
Emojis, *Hashtags*, Pronomes em 1ª pessoa	Objetiva	Empate
Caixa alta, Repetição de pontuação	Objetiva	***Pipeline***
Autoridades, Termos técnicos	Subjetiva	Empate
Erros de língua portuguesa	Subjetiva	***Pipeline***

cluiu erroneamente sinais sem apelo emocional intenso, como reticências. Além disso, o *pipeline* identificou um volume superior de abreviações informais e erros ortográficos.

As pistas de caráter subjetivo expuseram limitações em ambas as abordagens. Na identificação de autoridades, observou-se um empate com perfis complementares: o LLM foi mais eficaz em entidades institucionais formais (e.g., "TSE" e "forças armadas"), enquanto o *pipeline* destacou-se em referências informais e geográficas com uso político (e.g., "América Latina" e "nine"). A detecção de termos técnicos foi a mais complexa: o LLM demonstrou permissividade ao classificar palavras comuns como especializadas, ao passo que o *pipeline* gerou falsos positivos com gírias políticas.

5. Conclusão e trabalhos futuros

O *pipeline* de triagem inteligente apresentado demonstrou ser uma etapa de pré-processamento eficaz para mensurar objetivamente a utilidade de *tweets* para PLN, fundamentado em pistas linguísticas retóricas. A análise revelou que o emprego de uma arquitetura híbrida (unindo regras explícitas a interpretações contextuais) pode ser a abordagem mais promissora para a curadoria de textos informais.

Como trabalhos futuros, propõe-se a integração de recursos complexos, como vetores de palavras (*word embeddings*), especificamente o modelo Wang2Vec [Hartmann et al. 2017], para detecção semântica de fatos históricos, e o uso do *Toxic Language Dataset for Brazilian Portuguese* (ToLD-Br) [Leite et al. 2020] para treinar classificadores que aprimorem a medição do apelo emocional negativo, ampliando a cobertura do sistema. O código-fonte do pipeline encontra-se disponível publicamente em: `https://github.com/LALIC-UFSCar/ic-triagem-tweets-politica`.

Agradecimentos

O presente trabalho foi realizado com apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) via Programa Institucional de Bolsas de Iniciação Científica (PIBIC). Esta pesquisa integra o projeto "Análise de grandes volumes de dados políticos e redes complexas: mineração, modelagens e aplicações em Ciência Política Computacional" financiado pela FAPESP (#22/03090-0) e faz parte do Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação (INCT-TILD-IAR). Agradecemos também à Pró-Reitoria de Graduação (ProGrad) da UFSCar pelo auxílio financeiro concedido para a participação no evento. Em conformidade com as diretrizes do evento, os autores declaram o uso do modelo generativo de IA Gemini (Google) para auxiliar na revisão, formatação em LaTeX e aprimoramento da fluidez textual deste artigo. A responsabilidade integral pelo conteúdo, análises e conclusões permanece dos autores.

Referências

Al Sharou, K., Li, Z., and Specia, L. (2021). Towards a better understanding of noise in natural language processing. In Mitkov, R. and Angelova, G., editors, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 53–62, Held Online. INCOMA Ltd.

Albalak, A., Elazar, Y., Xie, S. M., Longpre, S., Lambert, N., Wang, X., Muennighoff, N., Hou, B., Pan, L., Jeong, H., Raffel, C., Chang, S., Hashimoto, T., and Wang, W. Y. (2024). A survey on data selection for language models.

Bowman, S. R. (2024). Eight Things to Know about Large Language Models. *Critical AI*, 2(2).

Caseli, H. d. M., Nunes, M. d. G. V., and Pagano, A. (2024). O que é PLN? In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, chapter 1. BPLN, 3 edition.

Chen, H. (2022). *Data Quality Evaluation and Improvement for Machine Learning*. PhD thesis, University of North Texas, USA.

Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., and Aluisio, S. (2017). Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks. In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, Uberlândia, Brazil.

Iasulaitis, S., Valejo, A. D. B., Greco, B. C., Perillo, V. G., Messias, G. H., Vicari, I., and Interfaces—Center for Sociopolitical Studies of Algorithms and Artificial Intelligence (2025). The Interfaces Twitter Elections Dataset: Construction process and characteristics of big social data during the 2022 presidential elections in Brazil. *PLOS ONE*, 20(2):e0316626.

Leal, S. E., Duran, M. S., Scarton, C. E., Hartmann, N. S., and Aluísio, S. M. (2024). NILC-Metrix: Assessing the Complexity of Written and Spoken Language in Brazilian Portuguese. *Language Resources and Evaluation*, 58(1):73–110.

Leite, J. A., Silva, D. F., Bontcheva, K., and Scarton, C. (2020). Toxic Language Detection in Social Media for Brazilian Portuguese: New Dataset and Multilingual Analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, Suzhou, China.

Lopes, L., Duran, M. S., Fernandes, P., and Pardo, T. A. S. (2022). PortiLexicon-UD: A Portuguese Lexical Resource According to Universal Dependencies Model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6635–6643, Marseille, France. European Language Resources Association.

Pinto, A., Gonçalo Oliveira, H., Figueira, Á., and Alves, A. O. (2017). Predicting the relevance of social media posts based on linguistic features and journalistic criteria. *New Generation Computing*, 35(4):451–472.

Silva, C. F. d. (2023). *Abordagem computacional para avaliação automática da qualidade da argumentação na dimensão retórica de tweets no domínio da política brasileira*. PhD thesis, Universidade Federal de São Carlos, São Carlos.

Wachsmuth, H., Naderi, N., Hou, Y., Bilu, Y., Prabhakaran, V., Thijm, T. A., Hirst, G., and Stein, B. (2017). Computational Argumentation Quality Assessment in Natural Language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Author Index

Ahad, Felipe R., 218
Alexandre, Dominick Maia, 1
Almada, Fabrycio Leite Nakano, 567
Alves, Felipe André Bach, 14
Aragão, Iago de Sousa, 163
Assis, Gabriel, 26, 209
Athaydes, Aline, 573, 600

Balreira, Dennis Giovani, 126
Barbosa, Antonio Ferreira de Castro, 540
Barbosa, Yuri de Almeida Malheiros, 113
Barnes, Yasmin Francisquetti, 562
Benetti, Ângelo Brandão, 621
Berndt, Gabriela, 41
Berton, Lilian, 150, 317
Bertotto, Eduarda, 297
Bezerra, Eduardo, 78
Binda, Mateus, 433
Boll, Heloísa Oss, 14
Borges, Luciana C. L. F., 348
Borges, Nathalia, 348
Brandão, Amanda Sammer do N., 50
Brito, Erick, 305

Campos, Andrew, 551
Cardoso, Luiza Rodrigues, 610
Cardoso, Paula Christina Figueira, 531, 616
Carpi, Miguel M., 62
Carpio, Paula Salem, 297
Carvalho, Cid Ivan C., 476
Carvalho, Flavio, 78
Carvalho, Rivânia da Paixão de Jesus, 454
Caseli, Helena, 270
Caseli, Helena de Medeiros, 605, 626
Cavalcante, Rerisson, 50, 405
Claro, Daniela Barreiro, 50, 270, 405, 573, 600
Comarim, Maria Julia Bernardo, 86
Coneglian, André V. Lopes, 200, 487, 499
Correa, João Victor M., 100
Corrêa, Pedro L. P., 258
Costa, Anna Helena Reali, 231
Cozman, Fabio G., 209

da Silva, Altigran Soares, 26, 331
da Silva, Carlos Vitor Cardoso, 616
da Silva, Gustavo Borba, 621
da Silva, Samuel Rios, 405, 600
Dantas, Larrissa, 270, 405
de Alencar, Leonel Figueiredo, 1
de Araújo, Isaac Sebastian Lima, 113
de Faria, Jefferson Felex, 567
de Freitas, Larissa Astrogildo, 551
de Lira, Humberto Nunes, 113
De Martino, José Mario, 594, 621
de Oliveira, Allan G., 218
de Oliveira, Brenda R. F., 578
de Sousa, Lilian Teixeira, 588
de Souza, Antônio Artur, 433
de Souza, Gabriel Z., 441
Di-Felippo, Ariani, 86, 139
dos Santos, Lucas, 361
dos Santos, Luiz G., 348
dos Santos, Valeria Vieira, 522
Drumond, Patricia Medyna Lauritzen de Lucena, 163
Duarte, Rômulo Henrique Nascimento, 113
Duran, Magali, 285, 499
Duran, Magali Sanches, 464
Dutra, Maykon Adriell, 567

Engler, Felix, 487
Evangelista, Guilherme, 420
Evangelista, Pedro Moreno Niella de Souza Sales, 546

Faé, Eduardo D., 126
Fernandes, Gerson Vitor P., 476
Ferri, Antonio Paulo Martins, 551
Figueiredo, Josiel M., 218
Finatto, Maria J. B., 297
Finatto, Maria Jose Bocorny, 175, 188
Finatto, Maria José, 405
Finger, Marcelo, 62, 245, 390, 562
Fortalesa, Gabriela, 556
França, Celso, 420
Freitas, Christian, 150
Freitas, Claudia, 270
Freitas, Kamylla Sejane Pouso, 567

Gamba, Federica, 285
Gamonal, Maucha Andrade, 487
Garcia, Gabriel L., 375
Gomes, Gabrielly Alves, 163
Gomes, Raquel Salcedo, 175, 188
Gonçalves, Eduardo Corrêa, 540
Gonçalves, Marcos, 420

Grandi, Roges Horacio, 175, 188, 297
Guedes, Gustavo, 78
Guimarães, Letícia Guedes, 200, 499

Hochgraf, Gustavo, 209

Jin, Mariana, 41
Jr., Wagner Meira, 420

Kempner, Thais R., 348

Leite, Guilherme Vieira, 594
Lima, Guilherme Dallmann, 551
Lima, Marjory S. S., 218
Lopes, Lucelene, 464
Lopes, Marcos, 512
Lotufo, Roberto, 305

Maia, Fabio Masaracchia, 231
Mancini, Gabriel Silvestre, 512
Mane, Babacar, 573, 600
Manesco, João Renato R., 375
Manssour, Isabel H., 441
Mariano, Kauan Divino Pouso, 567
Masid, Marina Ramalhete, 26
Mattos, Ely E. S., 487
Mauá, Denis D., 62
Mauá, Denis Deratani, 390
Mello, Guilherme L., 245
Miyaji, Renato O., 258
Monteiro, Victor Emanuel da Silva, 567
Moreira, Viviane Pereira, 126
Mota, Laila, 50
Mota, Lucas B. Bulcão, 270, 573

Nepomuceno, João, 26
Nunes, Eunice P. S., 348
Nunes, Maria das Graças Volpe, 464

Oliveira, Pedro Paulo Freire, 583

Paes, Aline, 26, 209, 270, 405
Pagano, Adriana, 41, 200, 487, 499
Pagano, Adriana S., 285, 433
Paiola, Pedro Henrique, 375
Papa, João Paulo, 375
Pappa, Gisele, 420
Pardo, Thiago Alexandre Salgueiro, 464
Pasqualini, Bianca, 175, 188, 297
Passos, Leandro A., 375
Pedrini, Hélio, 594
Peixoto, Jennifer Vitória da Silva, 567
Pereira, Jayr, 305
Pereira, Jonas Oliveira, 588
Pereira, Lucas, 305
Pereira, Mateus Araújo, 454
Pereira, Vivianny Khatly Medeiros, 113
Pinto, Guilherme Galvão de Oliveira, 594
Pires, Thiago Blanch, 583

Rabonato, Ricardo Trainotti, 317
Real, Livy, 26, 270, 331
Recamonde-Mendoza, Mariana, 14
Ribeiro, Leo Sampaio Ferraz, 610
Ribeiro, Silvana, 405
Rodrigues, Isabela C., 605
Roman, Norton Trevisan, 139
Rottschaefer, Eduardo, 26

Sales, Matheus, 348
Sanches, Felipe Bomfim, 50
Santos, Roney Lira de Sales, 361
Schuck, André da F., 375
Sena, Alfredo, 50
Serrano, Sofia, 41
Serras, Felipe Ribas, 390
Silva, Renato Moraes, 100
Silva, Vih A. S., 218
Silva, Vitoria Maria A., 476
Silveira, Igor C., 62
Silveira, Igor Cataneo, 390
Soares, Vanessa, 78
Soriano, Flávio, 420
Souza, Jackson Wilke da Cruz, 454, 531, 546, 556
Souza, João Pedro Holanda, 361
Souza, Marlo, 270, 588

Teixeira, Laura Pessine, 626

Vale, Oto Araújo, 100, 578
Vasconcelos, Allan Gabriel da Cunha, 113
Ventura, Thiago M., 218
Vianna, Daniela, 26
Villavicencio, Aline, 200

Wives, Leandro Krug, 175, 188, 297

Zeman, Daniel, 285