Socially Responsible and Explainable Automated Fact-Checking and Hate Speech Detection

Francielle Vargas
University of São Paulo
francielleavargas@usp.br

Fabrício Benevenuto
Federal University of Minas Gerais
fabricio@dcc.ufmg.br

Thiago A. S. Pardo
University of São Paulo
taspardo@icmc.usp.br

Abstract

This Ph.D. dissertation advances the state-of-the-art in Natural Language Processing (NLP) for Portuguese by proposing new and innovative data resources and explainable methods for hate speech detection and automated fact-checking. The thesis introduces several benchmark datasets for Brazilian Portuguese, HateBR, HateBRXplain, HateBRMoralXplain, MFTCXplain, MOL, and FactNews, which have been widely adopted by the research community and address critical gaps in the availability of high-quality annotated resources for Portuguese. In addition, this dissertation proposes novel post-hoc and self-explaining NLP methods: Sentence-Level Factual Reasoning (SELFAR), Social Stereotype Analysis (SSA), Contextual Bag-of-Words with Interpretable Input and Feature Optimization (B+M), Supervised Rational Attention (SRA), and Supervised Moral Rational Attention (SMRA). Across multiple tasks and datasets in Portuguese, these methods outperform baselines while improving interpretability and robustness, demonstrating that explainability and performance can be jointly optimized. Finally, this thesis has achieved significant national and international impact, being cited by leading universities and research institutes worldwide and fostering new M.Sc. and Ph.D. research projects in Brazil. Its scientific and social contributions have also been recognized with multiple prestigious national and international awards, including the Google LARA, the Maria Carolina Monard Best Thesis Award in Artificial Intelligence, the Trevisan Prize for Students "AI for Good" from Bocconi University for rigorous computer science research in AI with social impact, and the Diversity and Inclusion Award from the Association for Computational Linguistics (ACL). Lastly, this thesis has received two nominations for the Brazilian Computer Society Thesis Awards in Computer Science, and in Multimedia, Hypermedia, and Web.

1 Introduction

Although the proliferation of misinformation and hate speech is a global challenge, most existing fact-checking and hate speech detection models remain focused on English and rely on opaque architectures. As a result, these "black-box" approaches fail to provide meaningful rationales for their predictions, while Portuguese continues to lack high-quality corpora, benchmarks, and explainable methods. This lack of transparency introduces significant risks, including biased and unreliable model behavior, which has become a major concern in the Artificial Intelligence (AI) field (May et al., 2019). These limitations reveal a clear research gap, constraining both scientific progress and the development of transparent, effective, and culturally aware solutions for Portuguese language processing.

Hate speech detection technologies often inherit biases from their training data (Davidson et al., 2019), which may reflect subjective human annotations (Al Kuwatly et al., 2020; Sap et al., 2022; Vargas et al., 2022). These biases can reinforce social discrimination, such as racial and gender biases, especially when deployed at scale (Davani et al., 2023; Vargas et al., 2023a; Chuang et al., 2021; Sap et al., 2019; Davidson et al., 2019). Table 1 evidences this issue, showing two documents classified as hate speech by a fine-tuned BERT classifier (Kennedy et al., 2020). Note that while the second document from Gab social media contains explicit hate speech, the first document, extracted from the New York Times[1], does not. However, the fine-tuned BERT classifier incorrectly classified both as hate speech, due to the presence of group identifiers such as "black' and "Africans."

In similar settings, the issue of bias in fact-checking has been extensively analyzed in the literature (Park et al., 2021; Baly et al., 2018; Soprano et al., 2024; Vargas et al., 2023c). Biases

[1] https://www.nytimes.com/

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 35–42
April 13-16, 2026 ©2026 Association for Computational Linguistics

Documents	Predicted Class
For many Africans, the most threatening kind of ethnic hatred is black against black. - New York Times	hate speech
There is a great discrepancy between whites and blacks in SA. It is … [because] blacks will always be the most backward race in the world - Anonymous user, Gab.com	hate speech

Table 1: Two documents classified as hate speech by a fine-tuned BERT classifier. Group identifiers are underlined (Kennedy et al., 2020).

N.	Claims	Rate
1	"Latina workers make 54 cents for every dollar earned by white, non-Hispanic men" - Democratic Senator (tweet)	True
2	"A proposal in Syracuse would pay gang members $100-$200 per week to stay out of trouble" - Republican state legislator from New York (tweet)	Mostly True

Table 2: Examples of manually fact-checked claims and their assigned ratings published by PolitiFact.

(e.g., media bias, political bias,) in automated fact-checking may be also introduced during data training. Fact-checking models tend to rely on these biases without fully learning the underlying task. Instead, they often learn misleading correlations between news patterns and veracity labels as simplifications, rather than integrating the information to reason effectively (Wu et al., 2022). As a result, these models may not only fail when applied to real-life situations, where news patterns vary widely, but they can also undermine public trust and exacerbate political polarization (Kuzmin et al., 2020). For example, prior studies assessing the performance of human fact-checkers have reported conflicting findings (Amazeen, 2015) and identified significant inconsistencies among major fact-checking organizations such as PolitiFact, The Fact Checker[2], and FactCheck.org in their evaluations of statements on topics such as climate change, racism, and national debt (Marietta et al., 2015). In addition, only 10% of statements were fact-checked by both organizations in their study, with agreement primarily observed for statements classified as clearly true or false, but significantly lower agreement for ambiguous claims that highlight the inherent subjectivity in the manual assignment of veracity ratings, raising concerns about potential biases such as selective claim verification and inconsistencies in evaluation criteria (Nieminen and Rapeli, 2019). An example of this lack of precision in defining veracity ratings is shown in Table 2. Note that in the second example, the label "mostly true" is assigned despite the low consistence of the evidence.

Moreover, media and political biases can be introduced during data collection and training, influencing the behavior of fact-checking models. Consequently, rather than learning the fundamental task of factual verification, these models may inadvertently internalize misleading correlations between news patterns and veracity labels as shortcuts (Wu et al., 2022). This reliance on biased patterns not only compromises model performance in real-world scenarios, where news structures vary significantly, but also poses broader societal risks, including diminished public trust in fact-checking and increased political polarization (Kuzmin et al., 2020). Thus, ensuring the transparency and accountability of automated fact-checking systems is both a technical necessity and a social imperative. To address these challenges, fact-checking models should either incorporate post-hoc explanations for their outputs or embed interpretability mechanisms directly within their architecture (Kotonya and Toni, 2020).

Therefore, the inability of NLP models to provide rationales for their decisions remains a significant barrier to their broader adoption (Gongane et al., 2024). In the context of automated fact-checking and hate speech detection, this lack of transparency raises serious ethical concerns regarding model reliability and fairness. In response to these critical issues, this thesis advances transparent and explainable approaches for Portuguese language processing, with a particular focus on fact-checking and hate speech detection, two tasks that are central to combating misinformation and sustaining a fair and democratic society. Specifically, this Ph.D. thesis introduced several benchmark datasets for Portuguese (e.g., HateBR, HateBRXplain, HateBRMoralXplain, MFTCXplain, FactNews, and MOL), and developed new post-hoc and self-explaining computational methods (e.g., SELFAR, SSA, B+M, SRA, SMRA) to ensure that both data and models are explainable and socially

[2] https://www.washingtonpost.com/politics/fact-checker/

aligned. Notably, over multiple tasks in Portuguese, these methods reliably outperform the baselines and simultaneously improve interpretability and robustness, significantly contributing to advance the state-of-the-art in the computational processing of Portuguese.

2 Theoretical Background

Explainable Artificial Intelligence (XAI) systems can explain their reasoning to human users and express knowledge about how they will behave in the future (Guidotti et al., 2018; Adadi and Berrada, 2018). In this context, XAI methods provide the causes of a single prediction, a set of predictions, or all predictions of a model by identifying the input, model, or training data parts that most influence the model's outcome (Balkir et al., 2022). The key concept in explainability involves the types of explanations, often categorized in literature into two main groups: (i) local and global explanations and (ii) self and post-hoc explanations (Guidotti et al., 2018; Adadi and Berrada, 2018), as follow.

Local explanations: This first type of explanation provides information or justification for the model's prediction regarding a specific input. Furthermore, local explanations are also known as model-agnostic explanations, meaning they do not consider the structure of the model.

Global explanations: This second type of explanation arises directly from the prediction process. It provides justification by revealing how the model's predictive mechanism works, regardless of any specific input. Moreover, global explanations are also known as model-specific explanations, as they consider the internal structure of the model's process and rely on the specific architecture used.

In addition, explanations differ based on whether they are generated as part of the prediction process or require post-processing after the model makes a prediction. These are categorized as self-explaining and post-hoc explaining, as described below:

Self-explaining: This type of approach is also referred to as directly interpretable (Arya et al., 2019). It generates explanations simultaneously with predictions, utilizing information provided by the model during the prediction process. For example, decision trees and rule-based models exemplify global self-explaining models, whereas feature saliency approaches, such as attention mechanisms or feature engineering, serve as examples of local self-explaining models.

Post-hoc explaining: This approach generates explanations after the model has been built, requiring an additional operation performed after predictions are made. LIME (Local Interpretable Model-Agnostic Explanations) (Ribeiro et al., 2016) and SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017) are examples of post-hoc explaining methods. LIME provides local explanations for predictions by perturbing input data and observing the resulting changes in the model's predictions. In contrast, SHAP measures the contribution of each feature to the prediction by considering all possible combinations of features and can be used for both local and global explanations.

3 Problem, Motivation and Objectives

Advances in NLP models have led to a decline in transparency, increasing risks such as bias and reduced accountability, particularly in socially sensitive tasks such as fact-checking and hate speech detection. These challenges are particularly notable for Portuguese, which still lacks adequate resources and explainable methods. Hence, the main objectives of this thesis are: (i) to analyze the risks, and limitations of "black-box" NLP models applied to fact-checking and hate speech detection in Portuguese; (ii) to design and release benchmark datasets for Portuguese, enriched with expert annotations and human-interpretable rationales; (iii) to propose novel explainable NLP methods that integrate interpretability into the learning process while maintaining state-of-the-art predictive performance; and (iv) to contribute to the advancement of the Portuguese NLP community by providing open-access data, methods, and evaluation metrics.

4 Research Hypotheses

This thesis was guided by two main hypotheses. First, research on explainability is essential for building trust, ensuring fairness, and promoting the responsible use of AI, particularly for understanding the risks and social impacts of "black-box" NLP models for automated fact-checking and hate speech detection in Portuguese. Second, by revealing when models rely on spurious patterns or social biases, explainability supports error diagnosis, improves robustness, and enables accountability, which is crucial in hate speech detection and fact-checking, where understanding the reasons behind a decision is as important as the decision itself.

5 Research Methodology

The research adopts a data-driven and evaluation-oriented methodology, comprising: (i) the construction of benchmark datasets for Portuguese with expert annotations and rationale-level supervision; (ii) the development of explainable NLP models, including self-explaining and post-hoc architectures; (iii) the evaluation of models in Portuguese language using standard NLP metrics alongside explainability and fairness-oriented assessments; and (iv) a comparative analysis against strong state-of-the-art baselines to quantify advances in performance, transparency, and robustness.

6 Research and Social Impact on Portuguese Language Processing

The research presented in this dissertation has had a significant national and international impact on the computational processing of Portuguese, particularly in the context of socially-aware NLP tasks such as hate speech detection, misinformation, and explainable AI. The proposed methodologies and language resources directly address theoretical and technological challenges, with a special focus on Brazilian Portuguese, contributing to the advancement of robust, transparent, and socially responsible Portuguese-language NLP systems. A major outcome of this work is the creation of widely adopted benchmark datasets for Portuguese, including the **HateBR**,**HateBRXplain**, **HateBRMoralXplain**, **FactNews**, **Multilingual Offensive Lexicon (MOL)**, and **MFTCXplain** all of which have been extensively used for benchmarking, evaluation, and methodological development in NLP research. These resources fill critical gaps in high-quality, expert-annotated datasets for Portuguese. In the **international context**, the impact of this dissertation is evidenced by its adoption and citation by leading research institutions and universities such as Microsoft Research, Carnegie Mellon University, Harvard University, University of Maryland, University of Turin, Technical University of Munich, University of Bonn, University of Lisbon, National University of Singapore, Vrije Universiteit Amsterdam, and the Rochester Institute of Technology. These institutions have leveraged the proposed datasets, as well as our methods, such **B+M, SSA, SELFAR, SRA, SMRA**, for advancing multilingual and Portuguese-centered NLP research. Her international recognition continues to grow. In December 2025, **Dr. Vargas was invited to present her research at the highly prestigious ASML Synthesizer Open Showcase 2025**[3] **at Harvard University**, where she presented "*Brazil #WithoutHate: Self-Explaining and Moral-Aware AI for Hate Speech Detection*." This invited presentation highlights the maturity, originality, and societal relevance of her doctoral research, as well as its contribution to the development of explainable and fair AI systems for the processing of Portuguese. In the **national context**, this work has directly influenced graduate-level research in Brazil. Several universities, including UFMG, USP, UFF, UFCG, UDESC, UFOP, etc., have proposed or developed MSc and PhD theses centered on hate speech and misinformation analysis using the datasets and methodologies introduced in this dissertation. The relevance and visibility of this research also extend beyond publications. The author was invited to serve as a visiting researcher at the University of Southern California, and to present this work at Leibniz Institute for the Social Sciences. Additionally, she has actively contributed to international NLP research communities through service roles, including participation in the organizing committees of ICWSM (2021, 2022, 2023) and WOAH (2025) and DeepXplain (2025), as well as program committee in ACL, EMNLP, NAACL, LREC, and COLING. Overall, this dissertation contributes novel methods, datasets, and evaluation frameworks for Portuguese, promote reproducible research, and support the development of linguistically and socially responsible NLP technologies, fully aligning with the objectives of the PROPOR Best PhD Dissertation Award.

7 Thesis Outcomes

The thesis generated several outcomes, including:

1. **15 (fifteen) published papers in top-tier international conferences**: In total, 15 (fifteen) papers were published in top-tier AI and NLP international conferences, workshops, and journals (e.g., EMNLP, NAACL, LREC, RANLP, NLP journal, AAAI, etc.) (see list of publised papers).

2. **Several benchmark datasets for Portuguese**: HateBR[4] (Vargas et al., 2022), HateBRXplain (Salles et al., 2025), and HateBRMoralXplain (Vargas et al., 2026) (the first

[3] https://cyber.harvard.edu/events/asml-2025-synthesizer

[4] https://github.com/franciellevargas/HateBR

large-scale expert-annotated corpora for hate speech detection in Brazilian Portuguese, including hate speech and moral rationales for explainability); MOL[5] (Vargas et al., 2024a) (the first multilingual offensive lexicon, extracted from the HateBR containing 1,000 explicit and "clues" to identify implicit terms in Portuguese, translated into five languages while accounting for cultural aspects.); FactNews[6] (Vargas et al., 2023c) (a sentence-level annotated dataset for fact-checking in Portuguese); and MFTCXplain[7] (Trager et al., 2025), the first benchmark for evaluating moral reasoning of LLMs.

3. **5 (Five) new post-hoc and self-explaining computational methods evaluated for Portuguese**: SELFAR[8] (Vargas et al., 2024b) (the first explainable fact-checking method in Portuguese); SSA[9] (Vargas et al., 2023a) (a counter-stereotype post-hoc explanation method to assess social bias in hate speech classifiers); B+M[10] (Vargas et al., 2021) (a contextual BoW with interpretable input and feature optimization for explainable hate speech detection); and SRA (Eilertsen et al., 2025) and SMRA[11] (Vargas et al., 2026) (self-explaining methods that integrating hate speech and moral human-annotated rationales into deep learning models by attention alignment loss).

4. **A Web platform to combat hate speech in Portuguese**: NoHateBrazil[12] (Vargas et al., 2023b), a public platform for hate speech detection in Brazilian Portuguese.

5. **Microsoft leveraged the HateBR dataset to train LLMs**: Microsoft Research has used the proposed HateBR dataset to train two 2 (two) LLMs: CultureLLM (Li et al., 2024) and CulturePark (Li et al., 2025), demonstrating its clear relevance and practical applicability.

6. **National and international awards**: This work has been widely recognized through prestigious national and international awards and grants, including the Google LARA[13], the Maria Carolina Monard Award Best Thesis in AI[14], International Trevisan Prize for Students "AI for Good"[15], and was nominated for the Brazilian Computer Society's Thesis Awards in Computer Science (Vargas et al., 2025a) and Multimedia, Hypermedia and Web (Vargas et al., 2025b). Finally, the author of this thesis received a Diversity & Inclusion (D&I) award from the Association for Computational Linguistics (ACL) for EMNLP and NAACL 2024. This award is granted to Ph.D. students in recognition of their outstanding contributions and achievements in Natural Language Processing and Computational Linguistics, as well as to applicants from underrepresented groups presenting a paper at the main conference.

7. **Invited as an international visiting researcher**: The author of this thesis was invited to visit the University of Southern California (USC) and to present her work at an event of the Applied Social Media Lab at the *Berkman Klein Center for Internet & Society*, **Harvard University**.

8. **Invited as an international keynote speaker**: The author of this thesis served as a keynote speaker at the Conference on Harmful Online Communication at the Leibniz Institute for the Social Sciences (GESIS), in Germany, and at the Conference cum Conclave on Emerging Trends in Journalistic and Media Practices at DG Vaishnav College, in India.

9. **Invited roundtable lead at ICLR 2026 workshop**: The author of this thesis was invited to serve as a Roundtable Lead at the *Algorithmic Fairness Across Alignment Procedures and Agentic Systems (AFAA) Workshop*, held at the International Conference on Learning Representations 2026[16].

10. **Invited to top-tier NLP international conference and journals program committees**:

[5] https://github.com/franciellevargas/MOL
[6] https://github.com/franciellevargas/FactNews
[7] https://github.com/franciellevargas/MFTCXplain
[8] https://github.com/franciellevargas/SELFAR
[9] https://github.com/franciellevargas/SSA
[10] https://aclanthology.org/2021.ranlp-1.161/
[11] https://github.com/franciellevargas/SMRA
[12] http://143.107.183.175:14581/
[13] https://research.google/programs-and-events/phd-fellowship/
[14] https://www.icmc.usp.br/institucional/premios/premio-maria-carolina-monard
[15] https://cs.unibocconi.eu/news/trevisan-prize-students-ai-good-winners
[16] https://www.afciworkshop.org/afaa-2026

The author of this thesis was invited to serve on the program committees of Language Resources and Evaluation, Expert Systems with Applications, Online Social Networks and Media, Natural Language Processing Journals, as well as EMNLP, ACL, NAACL, RANLP, COLING, LREC, ICWSM, WWW and CIKM conferences.

11. **Co-organizer of top-tier international conferences and workshops**: The author of this thesis co-organized the prestigious ICWSM conference in 2021, 2022 and 2023, and was selected to organize ACL Workshop on Online Abuse and Harms (WOAH 2025) (Calabrese et al., 2025), and proposed the IJCNN 2025 Special Session on Explainable Deep Neural Networks for Responsible AI: Post-Hoc and Self-Explaining Approaches (DeepXplain 2025) [17].

12. **+10 new Ph.D. and M.Sc. theses, and undergraduate research projects in Brazil**: Several prestigious public universities in Brazil have used our resources to propose new Ph.D. and M.Sc. theses, including institutions such as USP, UFMG, UFF, UFOP, UFC, among others.

13. **Co-advisor of a computer science master's student with a published paper in a top-tier NLP venue**: The author of this thesis co-advised a master's student in computer science at DCC-UFMG, whose work was published at COLING 2025, a top-tier international NLP conference (see paper).

14. **High citations from prestigious international institutions**: Research by prestigious institutions (e.g. Carnegie Mellon University, University of Maryland, Harvard University), has been significantly influenced by the resources provided in this thesis, as reflected in a substantial number of citations (Google Scholar: 325 citations) [18].

15. **Collaboration with researchers across 10 countries on 5 continents**: The author of this thesis has led projects and collaborated with researchers in ten different countries on five continents (see research projects).

16. **Press Coverage**: The results of this dissertation received relevant media coverage, with nine reports, including one international outlet (see).

[17] https://deepxplain.github.io/

[18] https://tinyurl.com/2ckwn6vd

8 Ph.D. Dissertation

Link: https://tinyurl.com/7ps8ykys

9 Conclusion

This Ph.D. dissertation advances the state-of-the-art in Natural Language Processing for Portuguese by addressing the lack of high-quality resources and explainable methods for socially sensitive tasks such as hate speech detection and automated fact-checking. To bridge this gap, this thesis introduced several benchmark datasets for Brazilian Portuguese with expert annotations and rationale-level supervision. In addition, it proposed novel explainable computational methods that integrate post-hoc and self-explaining approaches to improve transparency in classical and neural machine learning models. Experimental results across multiple tasks demonstrate that explainability and predictive performance can be jointly optimized, leading to more robust and trustworthy AI systems. Beyond its methodological contributions, this research has had a significant scientific and societal impact. The datasets and methods proposed in this thesis have been widely adopted by the international research community, enabling new studies and graduate research projects while strengthening Portuguese-centered NLP. Overall, this dissertation contributes new resources and methods that advance explainable and socially responsible NLP for Portuguese, reinforcing the importance of transparency and accountability in modern AI and NLP systems.

Acknowledgements

The authors are grateful to São Paulo Research Foundation – FAPESP (grant #2025/01118-2) for financial support. This work had the support of 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. It was also supported by Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação (INCT-TILD-IAR).

References

Amina Adadi and Mohammed Berrada. 2018. A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160.

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, Held Online.

Michelle Amazeen. 2015. Revisiting the epistemology of fact-checking. *Critical Review*, 27(1):1–30.

Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilovic, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John T. Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. 2019. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *CoRR*, abs/1909.03012.

Esma Balkir, Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen Fraser. 2022. Challenges in applying explainability methods to improve the fairness of NLP models. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing*, pages 80–92, Seattle, USA.

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3528–3539, Brussels, Belgium.

Agostina Calabrese, Christine de Kock, Debora Nozza, Flor Miriam Plaza-del Arco, Zeerak Talat, and Francielle Vargas, editors. 2025. *Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH)*. Association for Computational Linguistics, Vienna, Austria.

Yung-Sung Chuang, Mingye Gao, Hongyin Luo, James Glass, Hung-yi Lee, Yun-Nung Chen, and Shang-Wen Li. 2021. Mitigating biases in toxic language detection through invariant rationalization. In *Proceedings of the 5th Workshop on Online Abuse and Harms*, pages 114–120, Held Online.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the 3rd Workshop on Abusive Language Online*, pages 25–35, Florence, Italy.

Brage Eilertsen, Røskva Bjørgfinsdóttir, Francielle Vargas, and Ali Ramezani-Kebrya. 2025. Aligning attention with human rationales for self-explaining hate speech detection. *Preprint*, arXiv:2511.07065.

Vaishali U. Gongane, Mousami V. Munot, and Alwin D. Anuse. 2024. A survey of explainable AI techniques for detection of fake news and hate speech on social media platforms. *Journal of Computational Social Science*, 7(1):587–623.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5).

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442, Held Online.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5430–5443, Barcelona, Spain.

Gleb Kuzmin, Daniil Larionov, Dina Pisarevskaya, and Ivan Smirnov. 2020. Fake news detection for the Russian language. In *Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media*, pages 45–57, Barcelona, Spain.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. Culturellm: Incorporating cultural differences into large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 84799–84838. Curran Associates, Inc.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2025. Culturepark: boosting cross-cultural understanding in large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Morgan Marietta, David C. Barker, and Todd Bowser. 2015. Fact-checking polarized politics: Does the fact-check industry provide consistent guidance on disputed realities? *The Forum*, 13(4):577–596.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 622–628, Minneapolis, Minnesota.

Sakari Nieminen and Lauri Rapeli. 2019. Fighting misperceptions and doubting journalists' objectivity: A review of fact-checking literature. *Political Studies Review*, 17(3):296–309.

Sungkyu Park, Jaimie Yejean Park, Jeong-han Kang, and Meeyoung Cha. 2021. The presence of unexpected biases in online fact-checking. *Harvard Kennedy School Misinformation Review*, 2(1).

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Isadora Salles, Francielle Vargas, and Fabrício Benevenuto. 2025. HateBRXplain: A benchmark dataset with human-annotated rationales for explainable hate speech detection in Brazilian Portuguese. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6659–6669, Abu Dhabi, UAE.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 5884–5906, Seattle, United States.

Michael Soprano, Kevin Roitero, David La Barbera, Davide Ceolin, Damiano Spina, Gianluca Demartini, and Stefano Mizzaro. 2024. Cognitive biases in fact-checking and their countermeasures: A review. *Inf. Process. Manage.*, 61(3).

Jackson Trager, Francielle Vargas, Diego Alves, Matteo Guida, Mikel K. Ngueajio, Ameeta Agrawal, Yalda Daryani, Farzan Karimi Malekabadi, and Flor Miriam Plaza-del Arco. 2025. MFTCXplain: A multilingual benchmark dataset for evaluating the moral reasoning of LLMs through multi-hop hate speech explanation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15709–15740, Suzhou, China.

Francielle Vargas, Isabelle Carvalho, Ali Hürriyetoğlu, Thiago Pardo, and Fabrício Benevenuto. 2023a. Socially responsible hate speech detection: Can classifiers reflect social stereotypes? In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1187–1196, Varna, Bulgaria.

Francielle Vargas, Isabelle Carvalho, Thiago A. S. Pardo, and Fabrício Benevenuto. 2024a. Context-aware and expert data resources for brazilian portuguese hate speech detection. *Natural Language Processing*, 31(2):435–456.

Francielle Vargas, Isabelle Carvalho, Fabiana Rodrigues de Góes, Thiago Pardo, and Fabrício Benevenuto. 2022. HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 7174–7183, Marseille, France.

Francielle Vargas, Isabelle Carvalho, Wolfgang Schmeisser-Nieto, Fabrício Benevenuto, and Thiago Pardo. 2023b. NoHateBrazil: A Brazilian Portuguese text offensiveness analysis system. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1180–1186, Varna, Bulgaria.

Francielle Vargas, Kokil Jaidka, Thiago Pardo, and Fabrício Benevenuto. 2023c. Predicting sentence-level factuality of news and bias of media outlets. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1197–1206, Varna, Bulgaria.

Francielle Vargas, Thiago Pardo, and Fabrício Benevenuto. 2025a. Socially responsible and explainable automated fact-checking and hate speech detection. In *Anais do XXXVIII Concurso de Teses e Dissertações*, pages 75–84, Porto Alegre, RS, Brasil. SBC.

Francielle Vargas, Thiago Pardo, and Fabrício Benevenuto. 2025b. Socially responsible and explainable automated fact-checking and hate speech detection. In *Anais Estendidos do XXXI Simpósio Brasileiro de Sistemas Multimídia e Web*, pages 25–26, Porto Alegre, RS, Brasil. SBC.

Francielle Vargas, Fabiana Rodrigues de Góes, Isabelle Carvalho, Fabrício Benevenuto, and Thiago Pardo. 2021. Contextual-lexicon approach for abusive language detection. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 1438–1447, Held Online.

Francielle Vargas, Isadora Salles, Diego Alves, Ameeta Agrawal, Thiago A. S. Pardo, and Fabrício Benevenuto. 2024b. Improving explainable fact-checking via sentence-level factual reasoning. In *Proceedings of the Seventh Fact Extraction and VERification Workshop*, pages 192–204, Miami, USA.

Francielle Vargas, Jackson Trager, Diego Alves, Surendrabikram Thapa, Matteo Guida, Berk Atil, Daryna Dementieva, Andrew Smart, and Ameeta Agrawal. 2026. Self-explaining hate speech detection with moral rationales. *Preprint*, arXiv:2601.03481.

Junfei Wu, Qiang Liu, Weizhi Xu, and Shu Wu. 2022. Bias mitigation for evidence-aware fake news detection by causal intervention. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2308–2313, New York, USA.