EVALITA 2023

The Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian

Proceedings of the Workshop

September 7-8, 2023

Copyright ©2023 for the individual papers by the papers' authors.
Copyright ©2023 for the volume as a collection by its editors.
This volume and its papers are published under the Creative Commons License Attribution 4.0 International (CC BY 4.0).

These proceedings have been published in the CEUR Workshop Proceedings series.
The original papers are available at: https://ceur-ws.org/Vol-3473.
The papers are mirrored in the ACL Anthology.

ISBN 979-12-5500-069-3

Table of Contents

Preface
Mirko Lai, Stefano Menini, Marco Polignano, Valentina Russo, Rachele Sprugnoli and Giulia Venturi 1

EVALITA 2023: Overview of the 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian
Mirko Lai, Stefano Menini, Marco Polignano, Valentina Russo, Rachele Sprugnoli and Giulia Venturi 3

Challenges for NLP shared tasks in the chatGPT era
Julio Gonzalo 10

ExtremITA at EVALITA 2023: Multi-Task Sustainable Scaling to Large Language Models at its Extreme
Claudiu D. Hromei, Danilo Croce, Valerio Basile and Roberto Basili 11

O-Dang at HODI and HaSpeeDe3: A Knowledge-Enhanced Approach to Homotransphobia and Hate Speech Detection in Italian
Chiara Di Bonaventura, Arianna Muti and Marco Antonio Stranisci 20

EMit at EVALITA 2023: Overview of the Categorical Emotion Detection in Italian Social Media Task
Oscar Araque, Simona Frenda, Rachele Sprugnoli, Debora Nozza and Viviana Patti 28

Emotion Hunters at EMit: Categorical Emotion Detection combining BERT and ChatGPT models (short paper)
Gianluca Calo', Francesco Massafra, Berardina De Carolis and Corrado Loglisci 36

App2Check at EMit: Large Language Models for Multilabel Emotion Classification (short paper)
Gioele Cageggi, Emanuele Di Rosa and Asia Uboldi 42

ABCD team at EMit: Ensemble Approach for Categorial Emotion Detection in Social Media Messages (short paper)
Dai Ba Nguyen, Nguyen Ngoc Phuong Uyen and Dang Van Thin 48

EmotivITA at EVALITA2023: Overview of the Dimensional and Multidimensional Emotion Analysis Task
Giovanni Gafà, Francesco Cutugno and Marco Venuti 53

ISTC-CNR at EmotivITA: Towards Better Dimensional and multi-dimensional Analysis of VAD Emotions (short paper)
Luana Bulla and Misael Mongiovì 61

PoliticIT at EVALITA 2023: Overview of the Political Ideology Detection in Italian Texts Task
Daniel Russo, Salud María Jiménez-Zafra, José Antonio García-Díaz, Tommaso Caselli, L. Alfonso Ureña-López and Rafael Valencia-García 65

INFOTEC-LaBD at PoliticIT: Political Ideology Detection in Italian Texts
Hiram Cabrera-Pineda, Eric Sadit Téllez and Sabino Miranda 73

UMUTeam at PoliticIT-EVALITA2023: Evaluating Transformer Model for Detecting Political Ideology in Italian Texts (short paper)
Ronghao Pan, Ángela Almela and Francisco García-Sánchez 79

Tübingen at PoliticIT: Exploring SVMs, Pretrained Language Models, and Linguistic Transfer for Ideology Detection in Social Media
Çağrı Çöltekin, Matteo Brivio and Fidan Can 85

URJC-Team at PoliticIT: Political Ideology Detection in Italian Texts Using Transformers Architectures (short paper)
Miguel Ángel Rodríguez-García 91

GeoLingIt at EVALITA 2023: Overview of the Geolocation of Linguistic Variation in Italy Task
Alan Ramponi and Camilla Casula 97

DANTE at GeoLingIt: Dialect-Aware Multi-Granularity Pre-training for Locating Tweets within Italy
Giuseppe Gallipoli, Moreno La Quatra, Daniele Rege Cambrin, Salvatore Greco and Luca Cagliero 105

Galliz at GeoLingIt: Enhancing BERT with Vocabulary Knowledge for Predicting the Region of Language Varieties of Italy
Tiziano Labruna and Simone Gallo 112

baptti at GeoLingIt: Beyond Boundaries, Enhancing Geolocation Prediction and Dialect Classification on Social Media in Italy
Alkis Koudounas, Flavio Giobergia, Irene Benedetto, Simone Monaco, Luca Cagliero, Daniele Apiletti and Elena Baralis 118

Salogni at GeoLingIt: Geolocalization by Fine-tuning BERT (short paper)
Ilaria Salogni 124

LangLearn at EVALITA 2023: Overview of the Language Learning Development Task
Chiara Alzetta, Dominique Brunato, Felice Dell'Orletta, Alessio Miaschi, Kenji Sagae, Claudia H. Sánchez-Gutiérrez and Giulia Venturi 130

BERT_4EVER at LangLearn: Language Development Assessment Model based on Sequential Information Attention Mechanism
Hongyan Wu, Nankai Lin, Shengyi Jiang and Lixian Xiao 138

IUSS-NeTS at LangLearn: The role of morphosyntactic features in language development assessment
Matilde Barbini, Emma Zanoli and Cristiano Chesi 144

bot.zen at LangLearn: regressing towards interpretability (short paper)
Egon W. Stemle, Martina Tebaldini, Francesca Bonanni, Filippo Pellegrino, Paolo Brasolin, Greta H. Franzini, Jennifer-Carmen Frey, Olga Lopopolo and Stefania Spina 150

HaSpeeDe3 at EVALITA 2023: Overview of the Political and Religious Hate Speech Detection task
Mirko Lai, Fabio Celli, Alan Ramponi, Sara Tonelli, Cristina Bosco and Viviana Patti 155

CHILab at HaSpeeDe3: Overview of the Taks A Textual (short paper)
Irene Siragusa and Roberto Pirrone 163

LMU at HaSpeeDe3: Multi-Dataset Training for Cross-Domain Hate Speech Detection
Viktor Hangya and Alexander Fraser 168

BERTicelli at HaSpeeDe 3: Fine-tuning and Cross-validating Large Language Models for Hate Speech Detection
Leonardo Grotti and Patrick Quick 174

HODI at EVALITA 2023: Overview of the first Shared Task on Homotransphobia Detection in Italian
Debora Nozza, Alessandra Teresa Cignarella, Greta Damo, Tommaso Caselli and Viviana Patti 180

CHILab at HODI: A minimalist approach (short paper)
Irene Siragusa and Roberto Pirrone 188

Team_Tamil at HODI: Few-Shot Learning for Detecting Homotransphobia in Italian Language
Rahul Ponnusamy, Prasanna Kumar Kumaresan, Kishore Kumar Ponnusamy, Charmathi Rajkumar, Ruba Priyadharshini and Bharathi Raja Chakravarthi 194

LCTs at HODI: Homotransphobic Speech Detection on Italian Tweets (short paper)
Davide Locatelli and Lorenzo Locatelli 201

DH-FBK at HODI: Multi-Task Learning with Classifier Ensemble Agreement, Oversampling and Synthetic Data
Elisa Leonardelli and Camilla Casula 206

MULTI-Fake-DetectiVE at EVALITA 2023: Overview of the MULTImodal Fake News Detection and VErification Task
Alessandro Bondielli, Pietro Dell'Oglio, Alessandro Lenci, Francesco Marcelloni, Lucia C. Passaro and Marco Sabbatini 213

HIJLI-JU-CLEF at MULTI-Fake-DetectiVE: Multimodal Fake News Detection Using Deep Learning Approach (short paper)
Sandip Sarkar, Nripen Tudu and Dipankar Das 221

AIMH at MULTI-Fake-DetectIVE: System Report (short paper)
Giovanni Puccetti and Andrea Esuli 226

PoliTo at MULTI-Fake-DetectiVE: Improving FND-CLIP for Multimodal Italian Fake News Detection
Lorenzo D'Amico, Davide Napolitano, Lorenzo Vaiani and Luca Cagliero 232

ACTI at EVALITA 2023: Automatic Conspiracy Theory Identification Task Overview (short paper)
Giuseppe Russo, Niklas Stoehr and Manoel Horta Ribeiro 238

Vitali at ACTI – Transformer-based Conspiracy Theory Identification
Michael Vitali, Vincenzo Scotti and Mark James Carman 244

UPB at ACTI: Detecting Conspiracies using fine tuned Sentence Transformers (short paper)
Andrei Paraschiv and Mihai Dascalu 251

sCambiaMenti at ACTI: Ensemble model with majority voting for Automatic Conspiracy Detection (short paper)
Selene Bianco and Daniela Salusso 257

Cicognini at ACTI: Analysis of techniques for conspiracies individuation in Italian (short paper)
Giacomo Cignoni and Alessandro Bucci 262

Giobergia at Multi-Task Transformer Tuning for Joint Conspiracy Theory Detection and Classification
Flavio Giobergia 268

NERMuD at EVALITA 2023: Overview of the Named-Entities Recognition on Multi-Domain Documents Task (short paper)
Alessio Palmero Aprosio and Teresa Paccosi 274

CLinkaRT at EVALITA 2023: Overview of the Task on Linking a Lab Result to its Test Event in the Clinical Domain
Begoña Altuna, Goutham Karunakaran, Alberto Lavelli, Bernardo Magnini, Manuela Speranza and Roberto Zanoli 279

Simple Ideas at CLinkaRT: LeaNER and MeaNER Relation Extraction
Marius Micluța-Câmpeanu and Liviu P. Dinu 287

Polimi at CLinkaRT: a Conditional Random Field vs a BERT-based approach
Vittorio Torri and Francesca Ieva 293

WiC-ITA at EVALITA2023: Overview of the EVALITA2023 Word-in-Context for ITAlian Task
Pierluigi Cassotti, Lucia Siciliani, Lucia C. Passaro, Maristella Gatto and Pierpaolo Basile . . 299

The Time-Embedding Travelers at WiC-ITA
Francesco Periti and Haim Dubossarsky 307

LG at WiC-ITA: Exploring the relation between semantic distance and equivalence in translation. (short paper)
Lorenzo Gregori 313

DisCoTex at EVALITA 2023: Overview of the Assessing DIScourse COherence in Italian TEXts task
Dominique Brunato, Davide Colla, Felice Dell'Orletta, Irene Dini, Daniele Paolo Radicioni and Andrea Amelio Ravelli 318

IUSSNets at DisCoTeX: A fine-tuned approach to coherence
Emma Zanoli, Matilde Barbini and Cristiano Chesi 326

MPG at DisCoTeX: Predicting Text Coherence by Tree-based Modelling of Linguistic Features (short paper)
Martina Galletti, Pietro Gravino and Giulio Prevedello 333

Preface to the EVALITA 2023 Proceedings

Mirko Lai[1], Stefano Menini[2], Marco Polignano[3], Valentina Russo[4], Rachele Sprugnoli[5] and Giulia Venturi[6]

[1]*University of Turin*
[2]*Fondazione Bruno Kessler*
[3]*University of Bari "Aldo Moro"*
[4]*Logogramma s.r.l.*
[5]*University of Parma*
[6]*Institute for Computational Linguistics "A. Zampolli" (CNR-ILC), Pisa*

1. Preface

EVALITA 2023 is an initiative of AILC (Associazione Italiana di Linguistica Computazionale)[1] and it is endorsed by the Italian Association for Artificial Intelligence (AIxIA)[2] and the Italian Association for Speech Sciences (AISV)[3]. As in the previous editions (http://www.evalita.it/), EVALITA 2023 is organized along a set of selected tasks, which provide participants with opportunities to discuss and explore both emerging and traditional areas of Natural Language Processing and Speech for Italian. The participation is encouraged for teams working both in academic institutions and industrial organizations.

This volume includes the reports of both task organizers and participants to all of the EVALITA 2023 challenges. In the 2023 edition, we coordinated the organization of 13 different tasks belonging to four research areas, being:

Affect

- **EMit** – Categorical Emotion Detection in Italian Social Media (O. Araque, S. Frenda, D. Nozza, V. Patti, R. Sprugnoli)
- **EmotivITA** – Dimensional and Multi-dimensional Emotion Analysis (G. Gafà, F. Cutugno, M. Venuti)

Authorship Analysis

- **PoliticIT** – Political Ideology Detection in Italian Texts (D. Russo, S.M. Jiménez-Zafra, J.A. García-Díaz, T. Caselli, M. Guerini, L.A. Ureña-López, R. Valencia-García)
- **GeoLingIt** – Geolocation of Linguistic Variation in Italy (A. Ramponi, C. Casula)
- **LangLearn** – Language Learning Development (C. Alzetta, D. Brunato, F. Dell'Orletta, A. Miaschi1 , K. Sagae, C.H. Sánchez-Gutiérrez, G. Venturi)

Computational Ethics

- **HaSpeeDe 3** – Political and Religious Hate Speech Detection (M. Lai, F. Celli, A. Ramponi, S. Tonelli, C. Bosco, V. Patti)
- **HODI** – Homotransphobia Detection in Italian (D. Nozza, G. Damo, A.T. Cignarella, T. Caselli, V. Patti)
- **MULTI-Fake-DetectiVE** – MULTImodal Fake News Detection and VErification (A. Bondielli, P. Dell'Oglio, A. Lenci, F. Marcelloni, L.C. Passaro)
- **ACTI** – Automatic Conspiracy Theory Identification (G. Russo, N. Stoehr, M. Horta Ribeiro)

New Challenges in Long-standing Tasks

- **NERMuD** - Named-Entities Recognition on Multi-Domain Documents (T. Paccosi, A. Palmero Aprosio)
- **CLinkaRT** – Linking a Lab Result to its Test Event in the Clinical Domain (B. Magnini, B. Altuna, A. Lavelli, M. Speranza, R. Zanoli)
- **WiC-ITA** – Word-in-Context task for Italian (P. Cassotti, L. Siciliani, L. Passaro, M. Gatto, P. Basile)
- **DisCoTEX** – Assessing DIScourse COherence in Italian TEXts (D. Brunato, D. Colla, F. Dell'Orletta, I. Dini, D.P. Radicioni, A.A. Ravelli)

The workshop, held in Parma on 7 and 8 September 2023, represented an occasion for all participants from both academic institutions and private companies to disseminate their work and results and to share ideas through

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
mirko.lai@unito.it (M. Lai); menini@fbk.eu (S. Menini); marco.polignano@uniba.it (M. Polignano); vrusso@logogramma.com (V. Russo); rachele.sprugnoli@unipr.it (R. Sprugnoli); giulia.venturi@ilc.cnr.it (G. Venturi)
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)
[1]https://www.ai-lc.it/en/
[2]https://aixia.it/en/
[3]https://www.aisv.it/

sessions dedicated to each task and a general discussion during the plenary event. We carried on with the tradition of the *Best system* award. It represented an incentive for students, IT developers, and researchers to push the boundaries of the state of the art by facing tasks in new ways, even if not winning.

We would like to thank our supporters: Talia[4], Almawave[5], Aptus.AI[6] and Logogramma SRL[7]. Our gratitude goes also to the University of Parma for hosting the event. In addition, we sincerely thank the *Best System* award committee (Felice Dell'Orletta, Bernardo Magnini, Azzurra Mancini, Stefano Menini, Viviana Patti) for providing their expertise and experience. Moreover, we acknowledge the AILC Board members for their trust and support. We warmly thank our invited speaker Julio Gonzalo, for having shared his knowledge and insights with his talk. Last but not least, we would like to thank all the task organizers and participants who made this edition special with their enthusiasm and creativity.

Chairs

Mirko Lai, University of Turin
Stefano Menini, Fondazione Bruno Kessler
Marco Polignano, University of Bari
Valentina Russo, Logogramma SRL
Rachele Sprugnoli, University of Parma
Giulia Venturi, Institute for Computational Linguistics "A. Zampolli", ILC-CNR

Steering Committee

Alan Ramponi, Fondazione Bruno Kessler
Alberto Lavelli, Fondazione Bruno Kessler
Alessandra Teresa Cignarella, University of Turin
Alessandro Bondielli, University of Pisa
Alessio Aprosio, Fondazione Bruno Kessler
Alessio Miaschi, Institute for Computational Linguistics "A. Zampolli", ILC-CNR
Andrea Amelio Ravelli, University of Bologna
Begoña Altuna, Universidad del País Vasco
Camilla Casula, University of Trento
Chiara Alzetta, Institute for Computational Linguistics "A. Zampolli", ILC-CNR
Cristina Bosco, University of Turin
Daniel Russo, Fondazione Bruno Kessler
Debora Nozza, Bocconi University
Dominique Brunato, Institute for Computational Linguistics "A. Zampolli", ILC-CNR
Fabio Celli, University of Trento
Felice Dell'Orletta, Institute for Computational Linguistics "A. Zampolli", ILC-CNR
Francesco Cutugno, University of Naples Federico II
Giovanni Gafà, University of Catania
Giuseppe Russo, Swiss Federal Institute of Technology
Greta Damo, Bocconi University
Irene Dini, University of Pisa
José García-Díaz, Universidad de Murcia
Kenji Sagae, University of California, Davis
Lucia Passaro, University of Pisa
Lucia Siciliani, University of Bari
Manuela Speranza, Fondazione Bruno Kessler
Marco Guerini, Fondazione Bruno Kessler
Maristella Gatto, University of Bari
Oscar Araque, Universidad Politécnica de Madrid
Pierluigi Cassotti, University of Bari
Pierpaolo Basile, University of Bari
Pietro Dell'Oglio, University of Florence
Rafael Valencia-García, Universidad de Murcia
Roberto Zanoli, Fondazione Bruno Kessler
Salud María Jiménez-Zafra, Universidad de Jaén
Sara Tonelli, Fondazione Bruno Kessler
Simona Frenda, University of Turin
Teresa Paccosi, University of Trento
Tommaso Caselli, University of Groningen
Viviana Patti, University of Turin

Best System Award Committee

Felice Dell'Orletta, Talia & Institute for Computational Linguistics "A. Zampolli", ILC-CNR
Stefano Menini, Fondazione Bruno Kessler
Viviana Patti, University of Turin
Azzurra Mancini, Logogramma SRL
Bernardo Magnini, Fondazione Bruno Kessler

Website

Rachele Sprugnoli, University of Parma
Sara Barcena, freelance designer

Booklet and badge

Carlo Maria Giordano, Logogramma SRL

Parma, September 2023

[4] https://talia.cloud/
[5] https://www.almawave.com/it/
[6] https://www.aptus.ai/
[7] https://www.logogramma.com/

EVALITA 2023: Overview of the 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian

Mirko Lai[1], Stefano Menini[2], Marco Polignano[3], Valentina Russo[4], Rachele Sprugnoli[5] and Giulia Venturi[6]

[1]*University of Turin*
[2]*Fondazione Bruno Kessler*
[3]*University of Bari "Aldo Moro"*
[4]*Logogramma s.r.l.*
[5]*University of Parma*
[6]*Institute for Computational Linguistics "A. Zampolli" (CNR-ILC), Pisa*

Abstract

EVALITA provides a shared framework for evaluating and comparing different Nautural Language Processing (NLP) and speech systems across various tasks suggested and organized by the Italian research community. These tasks represent scientific challenges and allow testing of methods, resources, and systems on shared benchmarks related to linguistic open issues and real-world applications, including considering multilingual and/or multi-modal perspectives. The EVALITA 2023 edition consisted of 13 different tasks grouped into four research areas: Affect, Authorship Analysis, Computational Ethics, and New Challenges in Long-standing Tasks. The participation saw 42 groups from 12 different countries, indicating an increasing international interest, partly due to the proposal of multilingual tasks. The final workshop showcases the results obtained and highlights the growing interest in using deep learning techniques based on Large Language Models as a new trend. Overall, EVALITA serves as a valuable platform for Italian and international researchers to explore NLP-related challenges, develop solutions, and foster discussions within the community.

Keywords

NLP, Evaluation, Italian, Speech, Tools

1. Introduction

The Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA) is the biennial initiative aimed at promoting the development of language and speech technologies for the Italian language. EVALITA is promoted by the Italian Association of Computational Linguistics (AILC)[1] and it is endorsed by the Italian Association for Artificial Intelligence (AIxIA)[2] and the Italian Association for Speech Sciences (AISV)[3].

EVALITA provides a shared framework where different systems and approaches can be scientifically evaluated and compared with each other with respect to a large variety of tasks, suggested and organized by the Italian research community. The proposed tasks represent scientific challenges where methods, resources, and systems can be tested against shared benchmarks representing linguistic open issues or real-world applications, possibly in a multilingual and/or multi-modal perspective. The collected datasets provide big opportunities for scientists to explore old and new problems concerning NLP in Italian as well as to develop solutions and discuss NLP-related issues within the community. Some tasks are traditionally present in the evaluation campaign, while others are completely new.

This paper introduces the tasks proposed at EVALITA 2023 and provides an overview of the participants and systems whose descriptions and obtained results are reported in these Proceedings. The EVALITA 2023 final workshop, held in Parma on September 7-8th, counts 13 different tasks. In particular, the selected tasks are grouped into four research areas (tracks) according to their objective and characteristics, namely: (i) *Affect*; (ii) *Authorship Analysis*; (iii) *Computational Ethics*; (iv) *New Challenges in Long-standing Tasks*.

This edition was participated by 42 groups whose members have affiliations in 12 different countries. The high number of tasks is in line with a clear trend towards an increasing volume of proposed tasks at EVALITA. In fact, we have witnessed a significant progression from the 5 tasks organized in the first EVALITA campaign in 2007 to a peak of 14 tasks in the latest 2020 edition. Although EVALITA is generally promoted and targeted to the Italian research community, this edition saw increas-

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
mirko.lai@unito.it (M. Lai); menini@fbk.eu (S. Menini); marco.polignano@uniba.it (M. Polignano); vrusso@logogramma.com (V. Russo); rachele.sprugnoli@unipr.it (R. Sprugnoli); giulia.venturi@ilc.cnr.it (G. Venturi)
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

[1]http://www.ai-lc.it
[2]http://www.aixia.it
[3]http://www.aisv.it

CEUR-WS.org/Vol-3473/overview.pdf

ing international participation, partly due to our strong encouragement for the submission of multilingual tasks. This confirms a general trend of internationalization for the campaign, which reached its maximum this year, as discussed further.

This overview is organized as follows: in Section 2 a brief description of the tasks belonging to the various areas is reported. Section 3 discusses the participation in the workshop referred to several aspects, from the research area to the affiliation of authors. Section 4 describes the criteria used to assign the best system across tasks award, made by an ad-hoc committee starting from the suggestions of task organizers and reviewers. Finally, section 5 points out both the obtained results and the future of the workshop.

2. EVALITA 2023 Tracks and Tasks

In the 2023 edition of EVALITA, 13 different tasks were proposed, peer-reviewed, and accepted. Data were produced by the task organizers and made available to the participants. For the future availability of this data, we are going to release them on GitHub[4], in accordance with the terms and conditions of the respective data sources. Such a repository will also reference alternative repositories managed by the task organizers. The tasks of EVALITA 2023 are grouped according to the following tracks corresponding to four broad research areas:

Affect

EMit - Categorical Emotion Detection in Italian Social Media [1]. It aims to provide the first evaluation framework for emotion detection in Italian texts at EVALITA, following the categorical approach and offering novel annotated data. It presents two subtasks: *i)* Subtask A, which consists of an emotion detection challenge, and *ii)* Subtask B, which introduces a novel problem of target detection of the expressed emotion.

EmotivITA - Dimensional and Multi-dimensional Emotion Analysis [2]. The first shared task for Italian that follows the dimensional approach in emotion analysis. It introduces a new Italian dataset annotated with the Valence, Arousal, and Dominance dimensions and has two subtasks: *i)* Dimensional emotion regression and *ii)* Multi-dimensional emotion regression.

Authorship Analysis

PoliticIT - Political Ideology Detection in Italian Texts [3]. It aims to extract politicians' ideology information from a set of tweets in Italian framed as a binary and a multiclass classification. The task is designed to be privacy-preserving and accompanied by a subtask targeting the identification of self-assigned gender as a demographic trait.

GeoLingIt - Geolocation of Linguistic Variation in Italy [4]. The first shared task on the geolocation from social media posts comprising content in language varieties other than standard Italian (i.e., regional Italian, and languages and dialects of Italy). It is articulated into two subtasks: *i)* coarse-grained geolocation, aiming at predicting the region in which the variety expressed in the post is spoken, and *ii)* fine-grained geolocation, aiming at predicting its exact coordinates.

LangLearn - Language Learning Development [5]. The first shared task on automatic language development assessment aimed at developing and evaluating systems to predict the evolution of the written language abilities of learners across several time intervals. It was conceived to be multilingual, relying on written productions of Italian and Spanish learners, and representative of L1 and L2 learning scenarios.

Computational Ethics

HaSpeeDe 3 - Political and Religious Hate Speech Detection [6]. The third edition of a shared task on the detection of hateful content in Italian tweets. Differently from the two previous editions (organized within EVALITA 2018 and 2020), it explores hate speech in strong polarised debates, concerning politics and religion. Participants are asked to predict hate speech in both in- and out-domain settings, using either only the provided textual content of the tweet or any kind of external data.

HODI - Homotransphobia Detection in Italian [7]. The first shared task for the automatic detection of homotransphobia in Italian. The challenge is organized into two subtasks: *i)* Subtask A focuses on the binary textual classification of homotransphobic tweets, *ii)* Subtask B is concerned with the identification of rationales for explainability in the form of textual spans of text.

MULTI-Fake-DetectiVE - MULTImodal Fake News Detection and VErification [8]. The first task on fake news detection in Italian that explores multimodality and wants to address the problem from two perspectives, represented by the two subtasks: *i)* sub-task 1 aimed to evaluate the effectiveness of multimodal fake news detection systems,

[4] https://github.com/evalita2023

ii) sub-task 2, which consists in gaining insights into the interplay between text and images. Both perspectives were framed as classification problems.

ACTI – Automatic Conspiracy Theory Identification [9]. The first shared task based exclusively on comments published on conspiratorial channels of telegram. It is articulated into two subtasks: *i)* Conspiratorial Content Classification consisting in identifying conspiratorial content and *ii)* Conspiratorial Category Classification about specific conspiracy theory classification.

New Challenges in Long-Standing Tasks

NERMuD - Named-Entities Recognition on Multi-Domain Documents [10]. It consists in extracting and classifying persons, organizations, and locations from documents in various domains. It is articulated into two subtasks: *i)* Domain-agnostic classification, where participants are required to identify and classify entities from different types of texts, i.e., news, fiction, and political speeches, using a single model, and *ii)* Domain-specific classification, where a different model can be used for each text type.

CLinkaRT - Linking a Lab Result to its Test Event in the Clinical Domain [11]. It is a relation extraction task based on clinical cases taken from the E3C corpus, i.e., Italian written documents reporting statements of clinical practice. The task consists in identifying test results and measurements and linking them to the textual mentions of the laboratory tests and measurements from which they were obtained.

WiC-ITA - Word-in-Context task for Italian [12]. The first shared task at EVALITA on determining if a word occurring in two different sentences has the same meaning or not. It has been modeled as both a binary classification and a ranking problem.

DisCoTEX - Assessing DIScourse COherence in Italian TEXts [13]. The first shared task focused on modeling discourse coherence for Italian real-word texts. It was organized into two independent tasks: a more traditional one, aimed at evaluating whether models are able to distinguish well-organized documents from corrupted ones, and a less explored one, which assesses the models' performance on texts evaluated for coherence by human raters.

3. Participation

EVALITA 2023 attracted the interest of a large number of researchers from academia and industry, for a total of 42 single teams composed of about 109 individuals participating in one or more of the 13 proposed tasks. After the evaluation period, **51** system descriptions were submitted (reported in these proceedings), i.e., a **12**% percentage decrease with respect to the previous EVALITA edition [14].

Moreover, task organizers allowed participants to submit more than one system result (called runs), for a total of 246 submitted runs. Table 1 shows the different tracks and tasks along with the number of participating teams and submitted runs. The data reported in the table is based on information provided by the task organizers at the end of the evaluation process. Such data represents an overestimation with respect to the systems described in the proceedings. The trends are similar, but there are differences due to groups participating in more than one task and groups that have not produced a system report.

Unlike previous EVALITA editions, the organizers were not discouraged from distinguishing the submissions between unconstrained and constrained runs[5]. In fact, some of them introduced subtasks based on external resources used for training, while others required both a constrained and an unconstrained run. Alternatively, they allowed participants the freedom to utilize external resources or augment the distributed datasets. This decision was motivated by the expectation that most participants would employ pre-trained Neural Language Models. Thus, the organizers wanted to assess the participants' creativity in adopting strategies beyond solely relying on these models.

Participation was quite imbalanced across different tracks and tasks, as reported in Figure 1: each rectangle represents a task whose size reflects the number of participants, while the color indicates the corresponding track.

In line with the past edition of EVALITA [14], the development of systems dedicated to identifying unethical behaviors or malicious intentions in texts, spanning various aspects of human society, remains a topic of significant interest to the community. In fact, as evidenced by the high participation, the shared tasks grouped under the "Computational Ethics" track obtained the most attention. However, for the first time, this year the second most participated track was the "Authorship Analysis" one, which is focused on analyzing text writing styles to capture diverse author characteristics. This is a quite new result since the same typology of track had a relatively

[5] A system is considered *constrained* when using the provided training data only; on the contrary, it is considered *unconstrained* when using additional material to augment the training dataset or to acquire additional resources.

Track	Task	Teams	Runs
Affect	EMit	4	8
	EmotivITA	2	5
Authorship Analysis	PoliticIT	7	8
	GeoLingIt	6	35
	LangLearn	5	18
Computational Ethics	HaSpeeDe 3	6	29
	HODI	8	22
	MULTI-Fake-DetectiVE	4	6
	ACTI	8	81
New Challenges in Long-standing Tasks	NERMuD	1	2
	CLinkaRT	3	6
	WiC-ITA	4	9
	DisCoTEX	3	19

Table 1
Number of participating teams and number of runs organized by track and task. The data reported is an overestimation with respect to the systems described in the proceedings (e.g. teams participating in more than a task are counted according to the number of tasks they participated in).

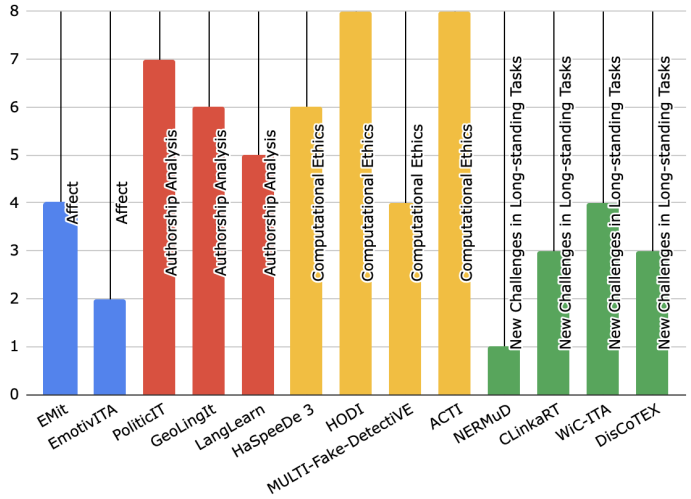


Figure 1: Number of participating teams organized by track (color) and task. The blue color is adopted for the track "*Affect*", the red color for "*Authorship Analysis*", yellow for "*Computational Ethics*", and green for "*New Challanges in Long-standing Tasks*".

low number of participants during the 2020 campaign. This shows the interest of the NLP community towards new and potentially more challenging areas of natural language understanding. It is worth noting that this year for the first time we introduced a new track solely dedicated to evaluating systems of emotions detection from two diverse perspectives (the "Affect" track). Additionally, we decided to keep the "New Challenges in Long-standing Tasks" track. Even if this track was among the least participated, the rationale behind this choice was to offer benchmarks for more conventional NLP tasks to the latest generation of Large Language Models. These models served as the foundation for the majority of the approaches devised by the participants, as illustrated in Section 5

It is worth noting that we also received a considerable number of tasks presented for the first time at EVALITA. Besides the two tasks centered around modeling different aspects of affect, namely *EMit* and *EmotivITA*, among them we can find *GeoLingIt*, *LangLearn*, *HODI*, and *ACTI*, which introduced novel problems. Interestingly, two of these newly introduced tasks received the highest num-

ber of submissions, showing the interest of the community in taking on new challenges.

In contrast to the 2020 edition, which saw a total of over 180 task organizers or participants, EVALITA 2023 experienced reduced participation. However, it is worth noting that the authorship of the 172 proceedings authors, including both participants and task organizers, reflects a greater diversity in terms of their origins, spanning 15 different countries. Notably, 70% of these contributors come from Italy, while the remaining 30% come from Institutions and companies abroad. The group of the 63 task organizers have affiliations in 6 countries (79% from Italy while 21% from Institutions and companies abroad). In summary, a noticeable increase was observed in the number of task organizers, particularly those affiliated with institutions abroad. In fact, the proportion of organizers with foreign affiliations more than doubled with respect to the previous edition, rising from 10% to 21% of the total organizers. This indicates a growing international interest in EVALITA. Notably, 6 out of the 13 tasks were organized by authors with mixed affiliations, combining both Italian and foreign institutions. This statistic aligns with one of the innovations we introduced this year. Indeed, during the call for tasks period, we encouraged the proposal of multilingual tasks, where participants were provided with datasets in both Italian and other languages. Up until now, only two tasks, namely *LangLearn* and *WiC-ITA*, provided participants with datasets in Italian and Spanish, and English, respectively. Although only a small number of organizers embraced this suggestion, we see it as a promising first step towards achieving a more international profile for EVALITA in the future.

As a last remark, we would like to notice that this year we had four teams that participated in multiple tasks. Among them, one team employed the same approach for two tasks (*HODI* and *HaSpeeDe 3*), while two other teams utilized distinct methods for two tasks each (*LangLearn* and *WiC-ITA*, and *LangLearn* and *DisCoTEX*). Particularly innovative was the approach taken by a single team, which submitted results for all 13 tasks, employing variations of the same model. In Section 5, we discuss how this feat was accomplished through the utilization of instruction-based models fine-tuned on all the EVALITA 2023 datasets using task-specific prompts.

4. Best System Across Tasks Award

In line with the previous edition, we confirmed the award to the best system across-task. The award was introduced with the aim of fostering student participation in the evaluation campaign and in the workshop.

A committee of 5 members (Felice Dell'Orletta, Bernardo Magnini, Azzurra Mancini, Stefano Menini, Viviana Patti) was asked to choose the best system across tasks. Three of the five members come from academia while two of them are from industry. The composition of the committee is balanced with respect to the level of seniority as well as to their academic background (computer science-oriented vs. humanities-oriented). In order to select a short list of candidates, the task organizers were invited to propose one candidate system participating in their tasks (not necessarily top-ranking). The committee was provided with the list of candidate systems and the criteria for eligibility, based on:

- *novelty* with respect to the state of the art;
- *originality*, in terms of identification of new linguistic resources, identification of linguistically motivated features, and implementation of a theoretical framework grounded in linguistics;
- *critical insight*, paving the way to future challenges (deep error analysis, discussion on the limits of the proposed system, discussion of the inherent challenges of the task);
- *technical soundness* and *methodological rigor*.

We collected **XX** system nominations from the organizers of **XX** tasks from across all tracks. The candidate systems are authored by 20 authors, among whom 12 are students, either at the master's or PhD level. The award recipient(s) will be announced during the final EVALITA workshop, during the plenary session, held online.

5. Final Remarks

The widespread adoption of Large Language Models (LLMs) was evident in the EVALITA 2023 challenge. LLMs, such as GPT-3 and its variants, have revolutionized the NLP landscape due to their ability to learn from large amounts of data and generate contextually relevant responses. These models have shown remarkable performance across various NLP tasks, and their usage was prominent in this edition of EVALITA. The confirmation of the massive use of LLMs underscores their effectiveness and potential in advancing NLP technology.

Traditional supervised learning approaches heavily rely on annotated data, which can be expensive and time-consuming to obtain. In response to this challenge, many participants in EVALITA 2023 proposed a semi-supervised approach using the prompting technique. The prompting technique involves providing the model with a few example inputs or a prompt to guide its response generation. This method allows leveraging limited labeled data while utilizing the model's language understanding capabilities to generalize to unseen instances. The adoption of the prompting technique showcases the interest in exploring more efficient and resourceful ways to tackle NLP tasks.

A noteworthy development in EVALITA 2023 was a team that participated in all tasks using the same approach, facilitated by prompt-based LLMs fine-tuning. While this approach showed promise, it also highlighted an essential observation: the performance of LLMs varies significantly across different NLP tasks. While LLMs are powerful models, they may not excel uniformly in all linguistic challenges. This underscores the need to understand the strengths and limitations of LLMs and to fine-tune them specifically for each task to achieve optimal results.

Another important outcome of the EVALITA 2023 challenge was the substantial increase in participation from groups outside Italy, making it one of the most attended editions by international teams. The rising international interest can be attributed to the growing significance of NLP and speech technologies on a global scale. The encouragement for multilingual tasks and the availability of shared datasets might have attracted researchers from different countries to participate actively. This trend signifies the growing impact and international recognition of the EVALITA initiative, facilitating collaboration and knowledge exchange among NLP communities worldwide.

To sum up, EVALITA 2023 outcomes demonstrate the dominance of LLMs in NLP, the exploration of semi-supervised approaches, the significance of task-specific fine-tuning, and the increasing internationalization of the initiative. These outcomes contribute to advancing the field of NLP, encouraging further research, and fostering a diverse and collaborative NLP community.

Acknowledgments

We would like to thank our sponsors: Talia[6], Almawave[7], APTUS.AI[8] and Logogramma[9]. Our gratitude goes also to the University of Parma for hosting the event. In addition, we sincerely thank the Best System award committee for providing their expertise and experience. Moreover, we acknowledge the AILC Board members for their trust and support. We warmly thank our invited speaker Julio Gonzalo, for having shared his knowledge and insights with his talk. Last but not least, we would like to thank all the task organizers and participants who made this edition special with their enthusiasm and creativity.

[6] https://talia.cloud/
[7] https://www.almawave.com/it/
[8] https://www.aptus.ai/
[9] https://www.logogramma.com/

References

[1] C. Alzetta, D. Brunato, F. Dell'Orletta, A. Miaschi, K. Sagae, C. H. Sánchez-Gutiérrez, G. Venturi, LangLearn at EVALITA 2023: Overview of the Language Learning Development Task, in: M. Lai, alii (Eds.), Proceedings of EVALITA 2023, CEUR.org, September 7th-8th 2023, Parma, 2023.

[2] G. Gafà, F. Cutugno, M. Venuti, EmotivITA at EVALITA 2023: Overview of the Dimensional and Multidimensional Emotion Analysis Task, in: M. Lai, alii (Eds.), Proceedings of EVALITA 2023, CEUR.org, September 7th-8th 2023, Parma, 2023.

[3] D. Russo, S. M. Jiménez Zafra, J. A. García-Díaz, T. Caselli, M. Guerini, L. A. Ureña López, R. Valencia-García, PoliticIT at EVALITA 2023: Overview of the Political Ideology Detection in Italian Texts Task, in: M. Lai, alii (Eds.), Proceedings of EVALITA 2023, CEUR.org, September 7th-8th 2023, Parma, 2023.

[4] A. Ramponi, C. Casula, GeoLingIt at EVALITA 2023: Overview of the Geolocation of Linguistic Variation in Italy Task, in: M. Lai, alii (Eds.), Proceedings of EVALITA 2023, CEUR.org, September 7th-8th 2023, Parma, 2023.

[5] O. Araque, S. Frenda, R. Sprugnoli, D. Nozza, V. Patti, EMit at EVALITA 2023: Overview of the Categorical Emotion Detection in Italian Social Media Task, in: M. Lai, alii (Eds.), Proceedings of EVALITA 2023, CEUR.org, September 7th-8th 2023, Parma, 2023.

[6] M. Lai, F. Celli, A. Ramponi, S. Tonelli, C. Bosco, V. Patti, HaSpeeDe3 at EVALITA 2023: Overview of the Political and Religious Hate Speech Detection task, in: M. Lai, alii (Eds.), Proceedings of EVALITA 2023, CEUR.org, September 7th-8th 2023, Parma, 2023.

[7] D. Nozza, A. T. Cignarella, G. Damo, T. Caselli, V. Patti, HODI at EVALITA 2023: Overview of the first Shared Task on Homotransphobia Detection in Italian, in: M. Lai, alii (Eds.), Proceedings of EVALITA 2023, CEUR.org, September 7th-8th 2023, Parma, 2023.

[8] A. Bondielli, P. Dell'Oglio, A. Lenci, F. Marcelloni, L. C. Passaro, M. Sabbatini, MULTI-Fake-DetectiVE at EVALITA 2023: Overview of the MULTImodal Fake News Detection and VErification Task, in: M. Lai, alii (Eds.), Proceedings of EVALITA 2023, CEUR.org, September 7th-8th 2023, Parma, 2023.

[9] G. Russo, N. Stoehr, M. Horta Ribeiro, ACTI at EVALITA 2023: Overview of the Conspiracy Theory Identification Task, in: M. Lai, alii (Eds.), Proceedings of EVALITA 2023, CEUR.org, September 7th-8th 2023, Parma, 2023.

[10] A. Palmero Aprosio, T. Paccosi, NERMuD at

EVALITA 2023: Overview of the Named-Entities Recognition on Multi-Domain Documents Task, in: M. Lai, alii (Eds.), Proceedings of EVALITA 2023, CEUR.org, September 7th-8th 2023, Parma, 2023.

[11] A. Begoña, G. Karunakaran, A. Lavelli, B. Magnini, M. Speranza, R. Zanoli, ClinkaRT at EVALITA 2023: Overview of the Task on Linking a Lab Result to its Test Event in the Clinical Domain, in: M. Lai, alii (Eds.), Proceedings of EVALITA 2023, CEUR.org, September 7th-8th 2023, Parma, 2023.

[12] P. Cassotti, L. Siciliani, L. C. Passaro, M. Gatto, P. Basile, WiC-ITA at EVALITA 2023: Overview of the EVALITA2023 Word-in-Context for ITAlian Task, in: M. Lai, alii (Eds.), Proceedings of EVALITA 2023, CEUR.org, September 7th-8th 2023, Parma, 2023.

[13] D. Brunato, D. Colla, F. Dell'Orletta, I. Dini, D. P. Radicioni, A. A. Ravelli, DisCoTEX at EVALITA 2023: Overview of the assessing DIScourse COherence in Italian TEXts task, in: M. Lai, alii (Eds.), Proceedings of EVALITA 2023, CEUR.org, September 7th-8th 2023, Parma, 2023.

[14] V. Basile, D. Croce, M. Di Maro, L. C. Passaro, EVALITA 2020: Overview of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), CEUR.org, Online, 2020.

Challenges for NLP shared tasks in the chatGPT era

Julio Gonzalo[1]

[1]Universidad Nacional de Educación a Distancia, Madrid, Spain

1. Abstract

The generative AI tsunami has turned Natural Language Processing upside down, and evaluation challenges (such as the ones conducted in SemEval, Evalita and IberLEF) are no exception. Since the introduction of chatGPT less than ten months ago, we have seen technical papers claiming that it annotates better than crowdworkers for many types of tasks, and at the same time that crowdworkers may be using chatGPT to complete their assignments in a substantial proportion. The best generative models are routinely passing exams from university degrees, but we have also learned that contamination is a key issue when evaluating large language models: models have probably seen the answers of the exams. Everything is fascinating and confusing.

In the talk we will discuss this situation, and we will review some evaluation challenges that are equally relevant in the post chaGPT era. We will pay special attention to evaluation metrics, learning with disagreement scenarios and the problem of measuring the relative performance of large language models between languages.

2. Short Biography

Julio Gonzalo is director of the Research Center in Natural Language Processing and Information Retrieval and deputy Vicerrector of Research at UNED (Madrid, Spain). Along his career he has worked on topics such as online reputation monitoring, Information Access technologies for Social Media, interactive cross-language search, toxicity and misinformation in Social Media, computational creativity and semantic similarity. He has also worked extensively in the design and assessment of evaluation metrics for a wide range of Artificial Intelligence problems, which led to a Google Faculty Research Award (together with Enrique Amigó and Stefano Mizzaro) for his work in this area. He has recently been co-chair of ACM SIGIR 2022 and co-funder and co-chair of IberLEF (2019-2022), the annual evaluation campaign for NLP systems in Spanish and other Iberian languages. He is currently leading an initiative of the Spanish government to measure the gap in Artificial Intelligence between Spanish and English.

Personal Website. https://sites.google.com/view/nlp-uned/people/julio-gonzalo

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT

julio@lsi.uned.es (J. Gonzalo)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/keynote.pdf

ExtremITA at EVALITA 2023: Multi-Task Sustainable Scaling to Large Language Models at its Extreme

Claudiu D. Hromei[1], Danilo Croce[1], Valerio Basile[2] and Roberto Basili[1]

[1]Università degli Studi di Roma Tor Vergata, Italy
[2]Università di Torino, Italy

Abstract

This paper explores the potential application of a monolithic neural model for all tasks in EVALITA 2023. We evaluated two models: `extremIT5`, an encoder-decoder model, and `extremITLLaMA` an instruction-tuned Decoder-only Large Language Model, specifically designed for handling Italian instructions. Our approach revolves around representing tasks in natural language, where we provide instructions to the model using prompts that define the expected responses. Remarkably, our best-performing model achieved first place in 41% of the subtasks and showcased top-three performance in 64%. These subtasks encompass various semantic dimensions, including Affect Detection, Authorship Analysis, Computational Ethics, Named Entity Recognition, Information Extraction, and Discourse Coherence.

1. Introduction

In recent years, Large Language Models (LLMs) have garnered substantial attention due to their remarkable performance across a wide range of NLP tasks. In addition to achieving state-of-the-art results in individual tasks, LLMs such as T5 [1], mT5 [2], IT5 [3], and FlanT5 [4] have demonstrated exceptional capabilities in solving various tasks individually and collectively through multi-task training paradigms.

In parallel, the generative power exhibited by models like GPT [5] and GPT3 [6], as well as the recent development of LLaMA [7] foundational models, has opened up new avenues for leveraging the concept of "prompting". This approach allows for modeling inductive tasks by formulating them linguistically, as natural language queries or instructions, enabling the model to provide accurate responses based on it. By combining the power of these models and the prompting technique, complex tasks can be addressed using a straightforward and intuitive interaction paradigm, without the need for task-specific feature engineering or neural architectures.

This paper presents `ExtremITA`, our approach developed for the EVALITA challenge [8]. The purpose of this work is to investigate how the adoption of a Large Language Model (LLM) can be taken to its extreme consequences by proposing a single model capable of tackling a wide array of heterogeneous tasks. Our methodology leverages an Encoder-Decoder model and a Decoder-only model, both trained on the union of all the available datasets for the challenge. By adopting a multi-task learning framework, our objective is to evaluate the applicability of a single model in effectively solving multiple tasks at once. Notably, our approach offers a significant advantage as it enables the resolution of diverse tasks by employing a unified architecture and fine-tuning based on input-output pairs. The EVALITA challenge serves as a robust testing ground for assessing the capabilities of LLMs across various Italian linguistic tasks, without any specific architectural requirement. Instead, we will trigger the model with task-specific prompts, such as "*Is there any mention of a conspiracy in this text? Answer yes or no.*" or "*How much consistent is this sentence, on a scale of 0 to 5?*".

The complete list of tasks in which the `ExtremITA` approach participated is here reported, in a wide range of semantic dimensions, including Affect Detection, Authorship Analysis, Computational Ethics, Named Entity Recognition, Information Extraction, and Discourse Coherence: *i)* EMit – Categorical Emotion Detection in Italian Social Media [9]; *ii)* EmotivITA – Dimensional and Multi-dimensional Emotion Analysis [10]; *iii)* PoliticIT – Political Ideology Detection in Italian Texts [11]; *iv)* GeoLingIt – Geolocation of Linguistic Variation in Italy [12]; *v)* LangLearn – Language Learning Development [13]; *vi)* HaSpeeDe 3 – Political and Religious Hate Speech Detection [14]; *vii)* HODI – Homotransphobia Detection in Italian [15]; *viii)* MULTI-Fake-DetectiVE – MULTImodal Fake News Detection and VErification [16]; *ix)* ACTI – Automatic Conspiracy Theory Identification [17, 18, 19]; *x)* NERMuD - Named-Entities Recognition on Multi-Domain Documents [20]; *xi)* CLinkaRT – Linking a Lab Result to its Test Event in the Clinical Domain [21]; *xii)* WiC-ITA – Word-in-Context task for Italian [22]; *xiii)* DisCoTEX – Assessing DIScourse COherence in Italian TEXts [23].

The aforementioned 13 tasks comprised 22 subtasks, where the proposed models ranked first in 9 subtasks (41%), and achieved a top-three position in 14 subtasks (64%). The adopted LLMs (especially LLaMA-based) proposed solution strongly supports the viability and high

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper13.pdf

performance of a single monolithic architecture, as it only requires modeling the tasks in natural language using prompts. This approach has been further reinforced by recent work [24], which indicates the same direction.

In the rest of the paper, Section 2 describes the adopted LLMs. Section 3 provides the results, accompanied by a brief error analysis. Finally Section 4 derives the conclusions.

2. Multi-task prompting in ExtremITA

The Transformer architecture [25] can be divided into two main components, each giving rise to distinct families of models. The encoder, exemplified by BERT [26], RoBERTa [27], and DeBERTa [28], is responsible for encoding input sequences and generating meaningful representations (embeddings) using the self-attention mechanism. On the other hand, the decoder, represented by models like GPT [5], GPT3 [6], and LLaMA [7], generates output sequences in an auto-regressive manner based on the input and previously generated output tokens. Additionally, another family of models, the Encoder-Decoder models, such as T5 [1] and BART [29], combine the strengths of both encoder and decoder components. These models maintain the integration of the two aforementioned blocks and they are usually used in tasks like machine translation, summarization, and question-answering, where complex input understanding as transduction is required.

A first effective application of an Encoder-Decoder architecture in a multi-task scenario is presented in [1]: in particular, the pre-training process of the so-called T5 involves training the model on a large corpus of diverse text data, which consists of a wide range of sources such as books, articles, and websites, but also texts involved in machine translation, classification and regression tasks. During pre-training, T5 utilizes a denoising objective, similar to other popular Transformer-based models like BERT and GPT. The model is trained to reconstruct masked or corrupted input text, which helps it learn meaningful representations and capture contextual information. One of the key strengths of T5 is its versatility. By casting various NLP tasks into a text-to-text format, it can be fine-tuned on a specific task simply by providing a prefix that serves as a description of the task and appropriate input-output pairs during fine-tuning. In practice, such an architecture can be triggered by concatenating the name of the task it is trained on with an input text, and it generates in output the expected solution to the task, e.g., a class label in a classification task or a text span that answers to a question. This flexibility eliminates the need for task-specific architectures or modifications, making it easier to apply T5 to different scenarios. Recently, this model was applied to hundreds of tasks in [24], while in [4] a systematic pre-training at large scale demonstrates the effectiveness within "zero-shot" or "few-shot" learning scenarios. In this paper, the first approach we adopted is based on T5, pre-trained on Italian texts, namely IT5 [3].

On the other hand, Decoder models are typically trained to be triggered by text, such as a natural language request or a piece of text intended for processing. These models generate text one word at a time, producing an output that can be an answer to a question or a solution to the given tasks or requests. Such models have the ability to essentially follow instructions, as exemplified by the recent release of ChatGPT. This characteristic holds a greater appeal, as tasks can be linguistically described using prompts, where the input sentence serves as contextual information. InstructGPT [30] is an extension of the GPT [6] language model explicitly designed to excel in multi-task scenarios when used with prompts. It combines the power of language models with the ability to follow instructions provided in the form of natural language prompts. Unlike conventional language models that generate text freely, InstructGPT is fine-tuned using human feedback to understand and generate text based on a given prompt and to select the best sequence that humans would have preferred. Another language model that adopts this instruction-tuning technique is Alpaca [31], which builds upon the LLaMA [7] foundational models. In the case of Alpaca, the authors created 175 sets of instructions, input sentences, and corresponding outputs. These were then used to generate variations using GPT 3.5, resulting in a collection of approximately $52,000$ instruction examples. The LLaMA model was further fine-tuned using this extensive dataset, a process referred to as instruction-tuning. The outcome of this effort was the Stanford Alpaca [31] as an instruction-following LLaMA model. More recently, an Italian counterpart called Camoscio [32] has overgone a similar intruction-tuning to Alpaca but on Italian data, essentially serving as the Italian equivalent. It is based on the same LLaMA model and it was instruction-tuned on the 52.000 instructions that were automatically translated into Italian using ChatGPT as in [32]. As the size of these models continues to grow, reaching trillions of parameters, there is a need for a way to fine-tune them effectively using modest GPU resources. The technique adopted in this paper is called Low-Rank Adaptation (LoRA [33]). LoRA involves freezing the weights of the pre-trained model and introducing trainable rank decomposition matrices into each layer of the Transformer architecture. This approach significantly reduces the number of trainable parameters for downstream tasks while avoiding additional inference latency.

To summarize, the ExtremITA approach for the EVALITA challenge focuses on efficiently modeling all

Task	Output Templates
EMit A	{"*Rabbia*", "*Anticipazione*", "*Disgusto*", "*Paura*", "*Gioia*", "*Amore*", "*Tristezza*", "*Sorpresa*", "*Fiducia*"}$^{+} \vee$ "*Neutrale*"
EMit B	{"*Direzione*", "*Argomento*", "*Entrambi*", "*Non specificato*"}
EmotivITA	"*Valenza: {0-5} Stimolo: {0-5} Controllo: {0-5}*"
PoliticIT	"*Gender*: {"*Uomo*", "*Donna*"} *PIB*: {"*Sinistra*", "*Destra*"} *PIM*: {"*Sinistra*", "*Destra*", "*Centro Sinistra*", "*Centro Destra*"}"
GeoLingIt	"*Regione*: {*Abruzzo*, .., *Veneto*} *Latitudine*: {} *Longitudine*: {}"
LangLearn*	{"*Corretto*", "*Non Corretto*"}
HaSpeeDe 3*	{"*Odio*", "*Non Odio*"}
HODI A*	{"*Omotransfobico*", "*Non Omotransfobico*"}
HODI B	<HOMOTRANSPHOBIA_MENTION>
MULTI-Fake -DetectiVE	{"*Certamente Falso*", "*Probabilmente Falso*", "*Probabilmente Vero*", "*Certamente Vero*"}
ACTI A*	{"*Cospirazione*", "*Non Cospirazione*"}
ACTI B	{"*Terrapiattista*", "*Covid*", "*Qanon*", "*Russia*"}
NERMuD	[<ENTITY_TYPE>] <TEXT_SPAN_THAT_EVOKES_ENTITY>
CLinkaRT	"[*BREL*] <RML_ENTITY_MENTION> [*SEP*] <EVENT_ENTITY_MENTION> [*EREL*]"
WiC-ITA*	{"*Uguale*", "*Differente*"}
DisCoTEX 1*	{"*Coerente*", "*Non Coerente*"}
DisCoTEX 2	{*0-5*}

Table 1
Output templates for ExtremITA models. In EMit A the model is requested to generate one or more labels from the first set ($^{+}$) or the text "*Neutrale*" if no emotion is expressed. In the tasks with * the extremITLLaMA model is requested to respond with {"*Si*", "*No*"}, for more details see Table 2.

available tasks using a single monolithic architecture, based on two independently tested models:

- extremIT5, An Encoder-Decoder model, based on IT5[1], consisting of approximately 110 million parameters. This model is trained by concatenating the name of the task and the input sentence/-paragraph in the input texts, each representing an example from a generic EVALITA task. Its purpose is to generate a piece of text that solves the target task.
- extremITLLaMA, an instruction-tuned Decoder-only model, built upon the LLaMA foundational models[2], with a total of 7 billion parameters. The initial model was trained using the LoRA technique on Italian translations[3] of Alpaca instruction data. This training enables the model to comprehend instructions in Italian. After training the adapters, they are merged into the original model to create an instruction-based model (using the "merge" procedure from [33]). Finally, this model is further fine-tuned using LoRA on instructions that reflect the EVALITA task. For each example from EVALITA, an input text is paired with a manually crafted question that simulates an instruction to be solved, accurately representing the specific task.

The next section describes how the 22 subtasks in EVALITA are encoded as prompts to fine-tune the above architectures.

Prompt Engineering in ExtremITA. The approach employed in this study draws inspiration from the original T5 and IT5 methodologies. Similar to those approaches, each training example is converted into a text-to-text format.

The model called extremIT5 is trained as a generic T5 model. In input, each example for an individual task is given to the neural architecture as concatenated after the task name. As an example, in the ACTI A task [17, 19] the input is just *"ACTI: Hanno votato tutti obbligo vaccinale, green pass, persecuzioni varie"*.

The output depends on the specific task. For a comprehensive compilation of outputs for the ExtremITA models, please refer to Table 1. In classification tasks involving only one label (such as EMit B, LangLearn, HaspeeDe 3, HODI A, MULTI-Fake-DetectiVE, ACTI A, ACTI B, WiC-ITA and DisCoTEX 1) the output is just the label of the target class. In the above example, the output would be "*Cospirazione*" as the input text reflects some conspiracy theory. In some tasks, such as PoliticIT [11], where a text is expected to be associated with the gender and the political inclination of the author, multiple labels reflecting such different dimensions are used, e.g., "*uomo sinistra centro-sinistra*". In EMit A [9] where multiple emotions can be triggered, these are provided as a sequence of labels. In regression tasks, such as EmotivITA [10] and DisCoTEX 2 [23], the output is the number to be

[1] https://huggingface.co/it5/it5-efficient-small-el32
[2] https://huggingface.co/decapoda-research/llama-7b-hf
[3] https://github.com/teelinsan/camoscio/tree/main/data

Task Name	Natural language instruction
EMit A	"*Quali emozioni sono espresse in questo testo? Puoi scegliere una o più emozioni tra 'rabbia', 'anticipazione', 'disgusto', 'paura', 'gioia', 'amore', 'tristezza', 'sorpresa', 'fiducia', o 'neutro'.*"
EMit B	"*Di cosa parla il testo, tra 'direzione', 'argomento', 'entrambi', 'non specificato'?*"
EmotivITA	"*Scrivi quanta valenza è espressa in questo testo su una scala da 1 a 5, seguito da quanto stimolo è espresso in questo testo su una scala da 1 a 5, seguito da quanto controllo è espresso in questo testo su una scala da 1 a 5.*"
PoliticIT	"*Scrivi se l'autore del testo è 'uomo' o 'donna', seguito dalla sua appartenenza politica tra 'destra', 'sinistra', 'centrodestra', 'centrosinistra'.*"
GeoLingIt	"*Scrivi la regione di appartenenza di chi ha scritto questo testo, seguito dalla latitudine, seguita dalla longitudine.*"
LangLearn	"*Questi due testi separati da [SEP] sono presentati nell'ordine in cui sono stati scritti? Rispondi sì o no.*"
HaSpeeDe 3	"*In questo testo si esprime odio? Rispondi sì o no.*"
HODI A	"*In questo testo si esprime odio omotransfobico? Rispondi sì o no.*"
HODI B	"*Con quali parole l'autore del testo precedente esprime odio omotransfobico? Separa le sequenze di parole con [gap].*"
MULTI-Fake -DetectiVE	"*L'evento riportato nel testo è 'certamente vero', 'probabilmente vero', 'probabilmente falso', o 'certamente falso'?*"
ACTI A	"*In questo testo si parla di una cospirazione? Rispondi sì o no.*"
ACTI B	"*Di quale teoria cospirazionista parla questo testo, tra 'Covid', 'Qanon', 'Terrapiattista', 'Russia'?*"
NERMuD	"*Scrivi le menzioni di entità nel testo, indicandone il tipo: [PER] (persona), [LOC] (luogo), [ORG] (organizzazione).*"
CLinkaRT	"*Trova i risultati dei test e delle misurazioni nel testo. Per ogni risultato, scrivi '[BREL]', seguito dal risultato seguito da '[SEP]', seguito dal test, seguito da '[EREL]'. Se non trovi nessun risultato, scrivi '[NOREL]'.*"
WiC-ITA	"*La parola compresa tra [TGTS] e [TGTE] ha lo stesso significato in entrambe le frasi? Rispondi sì o no.*"
DisCoTEX 1	"*Le due frasi precedenti, separate da '[SEP]', sono coerenti tra loro? Rispondi sì o no.*"
DisCoTEX 2	"*Quanto è coerente questa frase, su una scala da 0 a 5?*"

Table 2
List of the natural language instruction definition for all tasks for the `extremITLLaMA` model. Notice that these instructions have not been heavily optimized against individual tasks, also due to time constraints during the EVALITA challenge.

predicted within a specific range. In GeoLingIt [12], the models are requested to determine the region of origin of the tweet and the corresponding coordinates (latitude and longitude) based solely on the text. For instance, for the `extremIT5` model the task name ("GeoLingIt") is provided, while for `extremITLLaMA` a more in detail prompt is given: "*Scrivi la regione di appartenenza di chi ha scritto questo testo, seguito dalla latitudine, seguita dalla longitudine.*". For example, if the input sentence is "*Daje che je 'a famo!*", the model should provide the answer "*Lazio 41.8984164 12.54514535*", considering the use of the typical Roman dialect. This particular task combines both multi-label classification and regression, as it requires determining the region (classification) and providing the precise coordinates (regression) simultaneously. In HODI B [15] where the span of the offending text is expected to be extracted, it is simply provided as output. In NERMuD [20], the list of expected Named Entities is reported as a sequence of text spans, each associated with the corresponding entity type. CLinkaRT [21] focuses on extracting the names of medical tests performed on patients from an input text and linking them to the corresponding test results, treating it as a Relation Extraction problem. Here the relations are encoded with a slightly more complex form to summarize a list of relations, each associating an EVENT with a corresponding measure (or RML); as an example, the sentence "*CLinkaRT*: *Il PSA aumentava da 2 a 62 ng/ml.*" is associated with "[*BREL*] *2* [*SEP*] *PSA* [*EREL*] [*BREL*] *62 ng/ml* [*SEP*] *PSA* [*EREL*]" (where 2 and 62 reflect the RML while PSA is the test event).

In contrast, as `extremITLLaMA` is pre-trained to execute instructions, it leverages a structured prompt, which comprises the textual description of the task and the specification of the desired output format. For instance, when applied to the ACTI task, the instruction provided is "*In questo testo si parla di una cospirazione? Rispondi sì o no.*". The subsequent sentence to be evaluated is appended to this instruction. A comprehensive list of such instructions can be found in Table 2.

The decoder is thus expected to continue the sentence by generating the answer. In general, the same answers used in `extremIT5` are adopted. The only exception concerns the following binary classification tasks (LangLearn, HaSpeeDe 3, HODI A, ACTI A, WiC-ITA and DisCOTEX 1) where the instruction is only expected to answer *yes* or *no*, to reduce data sparseness.

3. Experimental Results

Experimental Setup. Models were trained using PyTorch, the Huggingface library and the Peft packages to implement the LoRA technique. Both models were trained on the unified dataset of all the tasks of EVALITA. Generally, one example in an EVALITA task corresponds to an example in our learning setting. Below are some exceptions. The dataset for the ACTI task was expanded by

incorporating some[4] sentences from dataset B and vice versa, resulting in an increase in the number of examples from 460 to 1,909 for ACTI A and from 300 to 777 for ACTI B.

Since in CLinkaRT only (long) documents were made available, these medical reports were segmented into smaller parts with a minimum of 50 characters and a maximum of 30 words using the Spacy library, respecting sentence boundaries. Moreover, we augmented this dataset with examples derived from the dataset made available in TESTLINK@IberLEF 2023[5] that contains medical reports in Spanish: although of a different language, these texts contain similar phenomena about events and measures that are generally language invariant and were useful to augment the dataset. This process significantly augmented the dataset, expanding it from 83 large documents to 3,903 shorter sentences. In general, this process recovered more than 95% of annotated relations. In the case of EMit, the dataset underwent a transformation where emoji representations were converted into textual descriptions, enhancing its compatibility with language models. GeoLingIt was modified to solve task A and task B simultaneously, enabling a single prediction for both tasks. For HODI B, only sentences expressing homotransphobia were considered, resulting in a reduction from 5,000 to 1,914 examples. The dataset of the LangLearn task was truncated into sentences with a maximum of 100 tokens, and additional examples with inverted sentence pairs (by flipping the label from positive to negative and vice versa), augmenting the dataset from 3,377 to 6,438 examples. In MULTI-Fake-DetectiVE we neglected images, and duplicate examples were removed (i.e., same text and different image), leading to a decrease from 1,058 to 860 examples. NERMuD was transformed into a sequence-to-sequence task from its original token classification format. In PoliticIT, each text was divided into sentences with a maximum length of 200 tokens, enabling more manageable input for language models. At classification time, a voting strategy was applied to select the final class about gender and political ideas, grouping all sentences written by the same author. Lastly, the WiC-ITA dataset was expanded by including examples[6] with inverted sentence pairs while preserving the same label, resulting in an increase from 5,610 to 6,600 examples. Overall, the entire dataset is composed of a total of 134,018 examples.

The `extremIT5` model underwent 10 epochs of training with a learning rate of $2 \cdot 10^{-5}$, while the `extremITLLaMA` model underwent 2 epochs of training with a learning rate of $3 \cdot 10^{-4}$. The models employed a batch size of 64 for `extremIT5` and 32 for `extremITLLaMA`. To optimize the models' performance, a linear scheduler with warmup was applied, utilizing a warmup ratio of 0.1. The `extremITLLaMA` model's training process utilized LoRA to refine the W_q, W_k, W_v and W_o modules of the transformer (for more details please refer to the original paper [33]), incorporating a matrix rank $R = 8$ and a parameter $\alpha = 16$ for the LoRA matrices. The decoding strategy in the generation phase used a beam search equal to 4, temperature of 0.2, with a top probability of 0.75 amongst the first 40 candidates. Two Tesla T4 GPUs with 16GB of memory each were used in parallel. This was particularly beneficial for the `extremITLLaMA` model, as its training duration exceeded 144 hours. The training data was divided into a 95% training set and a 5% validation set initially for hyper-parameter optimization. We release the source code on GitHub[7] for reproducing the experiment and dataset generation.

Results Discussion. The experimental results are reported in Table 3. We presented the tasks categorized by sub-task, followed by the Evaluation Metric, and the scores and ranks achieved by our `extremIT5` model, `extremITLLaMA` model, and the best competitor. The best-performing method for each subtask is highlighted in bold. Our systems, particularly `extremITLLaMA`, ranked first in 9 out of 22 subtasks (i.e., the 41% of subtasks) in EVALITA 2023. Additionally, it ranks in the top-three position in 14 subtasks, i.e., 64% of all tasks. However, we faced challenges in tasks such as GeoLingIt, LangLearn, and WiC-Ita, where our monolithic architectures demonstrated its limitations. These tasks specifically require a system to detect and analyze changes in the author's writing style or the contextual meaning of words. Our models are primarily designed for sentence classification or rewriting spans of input text to justify previous decisions (e.g., HODI).

There are also important considerations regarding the computational cost of both training and inference. Training `extremIT5` (made of "only" 110 million parameters) required approximately 12 hours on the entire EVALITA dataset, while `extremITLLaMA` (made of 7 billion parameters) took over 144 hours. In terms of inference, `extremITLLaMA` processes only 2 or 3 sentences per second, whereas `extremIT5` handles over almost one hundred sentences per second. This significant difference in processing speed makes the `extremITLLaMA` model less practical, despite its superior performance across a wide range of tasks. Additionally, the number of parameters between the two models differs by one order of magnitude, with `extremITLLaMA` having 7 billion parameters compared to `extremIT5`'s 110 million.

Overall, the above results are quite impressive, espe-

[4] Only the positive examples, i.e. the ones that involved any conspiracy theory, are added from the dataset A to B or viceversa.

[5] https://e3c.fbk.eu/testlinkiberlef

[6] Only the positive examples underwent sentence order flipping in order to rebalance the class distribution.

[7] https://github.com/crux82/ExtremITA

Task	SubTask	Eval metric	extremIT5		extremITLLaMA		Best Competitor	
			Score	R	Score	R	Score	R
Emit	A	F1	0.5086	2	**0.6028**	**1**	0.4994	3
	B	F1	0.6331	2	**0.6459**	**1**	0.6184	3
EmotivITA	B	Pears Val	0.7080		**0.8110**		0.8110	
		Pears Aro	0.4300	4	**0.6330**	**1**	0.6520	2
		Pears Dom	0.5480		**0.6300**		0.6540	
PoliticIT	-	F1	0.7034	7	0.7719	3	**0.8241**	**1**
GeoLingIt	A	F1	0.3999	10	0.3818	11	**0.6630**	**1**
	B	Avg Km	126.1	7	145.15	9	**97.74**	**1**
LangLearn	COWS	F1	0.1600	10	0.5500	8	**0.7500**	**1**
	CITA	F1	0.4100	10	0.6100	8	**0.9300**	**1**
HaSpeeDe 3	A	F1 - text.	0.9079	2	0.9034	3	**0.9128**	**1**
		F1 - context.	0.9079	2	0.9034	3	**0.9128**	**1**
	B	F1 - xRel.	0.5921	4	**0.6525**	**1**	0.6461	2
		F1 - xPolitic.	0.9079	2	0.9034	3	**0.9128**	**1**
HODI	A	F1	0.7431	10	0.7942	5	**0.8108**	**1**
	B	F1	0.6598	4	**0.7228**	**1**	0.7051	2
Multi-Fake-Detective	A	F1	na	na	0.5070	2	**0.5120**	**1**
	ATD	F1	0.3480	3	**0.4640**	**1**	0.4600	2
ACTI	A	F1	0.8183	7	0.8565	2	**0.8571**	**1**
	B	F1	0.8057	7	0.8556	5	**0.9123**	**1**
NERMUD	DAC	F1	0.8300	2	**0.8900**	**1**	na	3
CLinkaRT	-	F1	0.3382	4	0.5916	2	**0.6299**	**1**
Wic-Ita	A	F1 it-it	0.6100	5	0.5100	10	**0.7300**	**1**
		F1 it-en	0.6200	4	0.5400	8	**0.7400**	**1**
	B	F1 all	0.6100	5	0.5100	10	**0.7300**	**1**
DisCoTEX	1	Acc	0.7000	3	**0.8150**	**1**	0.7200	2
	2	HM*	0.0600	4	**0.6500**	**1**	0.6300	2

Table 3
ExtremITA ranks and results. Here each task is divided into the subtasks we participated in. Our models reported are `extremIT5` and `extremITLLaMA` and as a comparison the best competitor (either that won or placed higher in the ranking). In bold the rank and the scores of the winning systems. The HM* measure for the Discotex task refers to the Harmonic Mean between Pearson's and Spearman's correlations.

cially when considering that no task-specific architectural designs were applied. Instead, a single LLM was utilized, demonstrating competitive performance across almost all tasks. The key to achieving such results seems to lie in properly prompting the model with natural language requests or employing task-specific encoding techniques for the outputs. We can expect higher results to be achieved using larger LLMs such as LLaMA 65B. To conduct a more comprehensive evaluation and optimization, it would have been beneficial to explore a broader range of architectures and thoroughly investigate all the hyper-parameters of the models. The estimation of these parameters was done hastily due to the time constraints imposed by the EVALITA deadlines and the extensive commitment required for the parallel completion of all 13 tasks.

Error Analysis. Since our team participated in all the tasks, it would be unfeasible to provide a deeper analysis of each individual result in this report. However, in order to gain some insight into the inner working of the two models we employed, here we present some error analysis carried out on two tasks. We selected a task where our systems ranked very high, and one where they ranked very low. In the EmIt task A, `extremITLLaMA` ranked first in the official ranking, and `extremIT5` was second. The task is a multi-label classification problem, where the labels are eight emotions defined by Plutchik [34] plus "love" and a label for neutral texts. Table 4 reports the performance of the two `ExtremITA` systems broken down by labels. It is interesting to notice that the advantage shown by `extremITLLaMA` on the aggregated result comes from a skewed distribution over the labels. In particular, `extremIT5` is hardly capable of modeling Fear, which is also the least represented label in the test set. An inverse correlation between the number of positive instances in the test set and the gain in performance of `extremITLLaMA` with respect to `extremIT5` is indeed present. This indicates that `extremITLLaMA` is better suited than `extremIT5` for the classification of sparser phenomena. Moreover, `extremITLLaMA` shows superior capability in modeling and correctly predicting every emotion, besides "Trust", where `extremIT5` results in a better performance.

In the LangLearn task, our systems ranked quite low, respectively 8th place for `extremITLLaMA` and 10th place for `extremIT5`. LangLearn is a text pair classi-

Label	extremIT5			extremITLLaMA			Δ			Support
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	
Anger	0.500	0.464	0.481	0.759	0.393	0.518	0.259	-0.071	0.037	56
Anticipation	0.690	0.471	0.559	0.675	0.612	0.642	-0.015	0.141	0.083	85
Disgust	0.554	0.594	0.573	0.674	0.588	0.628	0.120	-0.006	0.055	165
Fear	1.000	0.077	0.143	0.636	0.538	0.583	-0.364	0.461	0.440	13
Joy	0.684	0.520	0.591	0.648	0.590	0.618	-0.036	0.070	0.027	100
Love	0.708	0.330	0.450	0.745	0.398	0.519	0.037	0.068	0.069	103
Neutral	0.705	0.614	0.656	0.657	0.757	0.704	-0.048	0.143	0.047	210
Sadness	0.584	0.474	0.523	0.750	0.537	0.626	0.166	0.063	0.103	95
Surprise	0.344	0.539	0.420	0.632	0.422	0.506	0.288	-0.117	0.086	102
Trust	0.679	0.699	0.688	0.698	0.673	0.685	0.019	-0.026	-0.003	272

Table 4
Performance in terms of Precision, Recall and F1-measure of our systems on the EmIt A task, where the Δ column is the difference between `extremITLLaMA` and `extremIT5`.

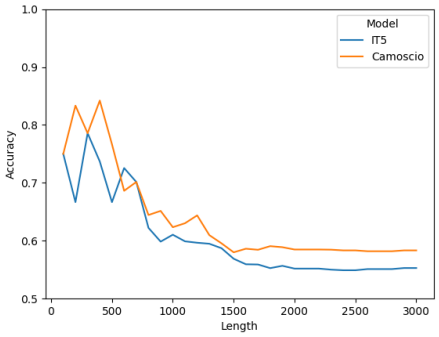


Figure 1: Accuracy of our systems on the LangLearn test set, with texts removed that are longer than an increasing threshold (horizontal axis).

fication task where the most informative features are expected to be stylistic, rather than semantic, to capture the development in language learning of the author of the texts. With this premise, we were anticipating a subpar performance by our transformer-based models from the beginning. However, another relevant characteristic of this task is the length of the texts. For computational reasons, we had to cut the texts to 100 tokens or less, therefore leaving out a significant portion of the data — we retained exactly 24.6% of the tokens from the two training sets combined. We checked the impact of the text size on the accuracy of the prediction, under the hypothesis that longer texts in the test set (which were cut by our systems to a greater extent) are penalized. The plot in Figure 1 shows the accuracy of our systems against portions of the test set where the texts were filtered by size. The number on the horizontal axis is a threshold on the minimum size in terms of characters of the two texts forming an instance of the test set. Indeed, the downward trend indicates that the predictions of our systems are more accurate on shorter pairs of texts, while more and more errors are made by both systems on longer texts.

4. Conclusions

In a recent position paper with a provocative title, Basile [35] asks himself "is EVALITA done?", referring to the mounting trend of LLMs and zero-shot approaches in NLP and their impact on the evaluation campaign. Judging by the results presented in this report, the answer is still the same as the original paper, i.e., *no*. The variety and challenge offered by the tasks of EVALITA continue to represent a fundamental resource to understand and develop language resources and tools for the Italian language, as shown, for instance, by the variability of the ranking obtained by our transformer-based models. However, the raw performance of `extremIT5` and `extremITLLaMA`, with minimal adaptations and tuning, is undoubtedly pushing the limits of some tasks, especially text classification tasks with roots in text semantics. In any case, these results once again confirm the huge potential of LLMs and their applicability in real-world scenarios. It is important to note that this experiment, while not conclusive, used the smallest available models due to their size limitations. Additionally, it would be worthwhile, from a sustainability standpoint, to explore the results that can be achieved by significantly reducing the amount of annotated data available through zero or few-shot learning approaches.

Acknowledgments

We would like to thank the "Istituto di Analisi dei Sistemi ed Informatica - Antonio Ruberti" (IASI) for supporting the experimentations. Claudiu Daniel Hromei is a Ph.D. student enrolled in the National Ph.D. in Artificial Intelligence, XXXVII cycle, course on *Health and life sciences*, organized by the Università Campus Bio-Medico di Roma.

References

[1] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (2020) 140:1–140:67. URL: http://jmlr.org/papers/v21/20-074.html.

[2] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mt5: A massively multilingual pre-trained text-to-text transformer, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the NAACL-HLT 2021, Online, June 6-11, 2021, Association for Computational Linguistics, 2021, pp. 483–498. URL: https://doi.org/10.18653/v1/2021.naacl-main.41. doi:10.18653/v1/2021.naacl-main.41.

[3] G. Sarti, M. Nissim, IT5: large-scale text-to-text pretraining for italian language understanding and generation, CoRR abs/2203.03759 (2022). URL: https://doi.org/10.48550/arXiv.2203.03759. doi:10.48550/arXiv.2203.03759. arXiv:2203.03759.

[4] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, S. Narang, G. Mishra, A. Yu, V. Y. Zhao, Y. Huang, A. M. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, J. Wei, Scaling instruction-finetuned language models, CoRR abs/2210.11416 (2022). URL: https://doi.org/10.48550/arXiv.2210.11416. doi:10.48550/arXiv.2210.11416. arXiv:2210.11416.

[5] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving language understanding by generative pre-training (2018).

[6] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, CoRR abs/2005.14165 (2020). URL: https://arxiv.org/abs/2005.14165. arXiv:2005.14165.

[7] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. arXiv:2302.13971.

[8] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[9] O. Araque, S. Frenda, R. Sprugnoli, D. Nozza, V. Patti, EMit at EVALITA 2023: Overview of the Categorical Emotion Detection in Italian Social Media Task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[10] G. Gafà, F. Cutugno, M. Venuti, Emotivita at EVALITA2023: Overview of the dimensional and multidimensional emotion analysis task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[11] D. Russo, S. M. Jiménez-Zafra, J. A. García-Díaz, T. Caselli, M. Guerini, L. A. Ureña-López, R. Valencia-García, Overview of PoliticIT2023@EVALITA: Political Ideology Detection in Italian Texts, 2023.

[12] A. Ramponi, C. Casula, GeoLingIt at EVALITA 2023: Overview of the task on geolocation of linguistic variation in Italy, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[13] C. Alzetta, D. Brunato, F. Delll'Orletta, A. Miaschi, K. Sagae, C. H. Sánchez-Gutiérrez, G. Venturi, Langlearn at evalita 2023: Overview of the language learning development task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[14] M. Lai, F. Celli, A. Ramponi, S. Tonelli, C. Bosco, V. Patti, Haspeede3 at evalita 2023: Overview of the political and religious hate speech detection task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[15] D. Nozza, A. T. Cignarella, G. Damo, T. Caselli, V. Patti, HODI at EVALITA 2023: Overview of the Homotransphobia Detection in Italian Task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[16] A. Bondielli, P. Dell'Oglio, A. Lenci, F. Marcelloni, L. C. Passaro, M. Sabbatini, Multi-fake-detective at evalita 2023: Overview of the multimodal fake news detection and verification task, in: Proceed-

ings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[17] P. Russo, N. Stoehr, M. Horta Ribeiro, Subtask a- conspiratorial content classification, 2023. URL: https://kaggle.com/competitions/acti-subtask-a.

[18] P. Russo, N. Stoehr, M. Horta Ribeiro, Subtask b - conspiracy category classification, 2023. URL: https://kaggle.com/competitions/acti-subtask-b.

[19] G. Russo, L. Verginer, M. H. Ribeiro, G. Casiraghi, Spillover of antisocial behavior from fringe platforms: The unintended consequences of community banning, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 17, 2023, pp. 742–753.

[20] A. Palmero Aprosio, T. Paccosi, Nermud at evalita 2023: Overview of the named-entities recognition on multi-domain documents task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[21] B. Altuna, G. Karunakaran, A. Lavelli, B. Magnini, M. Speranza, R. Zanoli, CLinkaRT at EVALITA 2023: Overview of the Task on Linking a Lab Result to its Test Event in the Clinical Domain , in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[22] P. Cassotti, L. Siciliani, L. Passaro, M. Gatto, P. Basile, Wic-ita at evalita2023: Overview of the evalita2023 word-in-context for italian task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[23] D. Brunato, D. Colla, F. Dell'Orletta, I. Dini, D. P. Radicioni, A. A. Ravelli, Discotex at evalita 2023: Overview of the assessing discourse coherence in italian texts task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[24] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, J. Wei, Scaling instruction-finetuned language models, 2022. arXiv:2210.11416.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, CoRR abs/1706.03762 (2017). URL: http://arxiv.org/abs/1706.03762. arXiv:1706.03762.

[26] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the NAACL 2019, 2019, pp. 4171–4186. URL: https://doi.org/10.18653/v1/n19-1423. doi:10.18653/v1/n19-1423.

[27] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[28] P. He, X. Liu, J. Gao, W. Chen, Deberta: decoding-enhanced bert with disentangled attention, in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, 2021. URL: https://openreview.net/forum?id=XPZIaotutsD.

[29] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, CoRR abs/1910.13461 (2019). URL: http://arxiv.org/abs/1910.13461. arXiv:1910.13461.

[30] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, R. Lowe, Training language models to follow instructions with human feedback, 2022. arXiv:2203.02155.

[31] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T. B. Hashimoto, Stanford alpaca: An instruction-following llama model, https://github.com/tatsu-lab/stanford_alpaca, 2023.

[32] A. Santilli, Camoscio: An italian instruction-tuned llama, https://github.com/teelinsan/camoscio, 2023.

[33] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, W. Chen, Lora: Low-rank adaptation of large language models, CoRR abs/2106.09685 (2021). URL: https://arxiv.org/abs/2106.09685. arXiv:2106.09685.

[34] R. Plutchik, H. Kellerman, Theories of emotion 1 (1980).

[35] V. Basile, Is EVALITA done? on the impact of prompting on the italian NLP evaluation campaign, in: D. Nozza, L. C. Passaro, M. Polignano (Eds.), Proceedings of the Sixth Workshop on Natural Language for Artificial Intelligence (NL4AI 2022), volume 3287 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 127–140. URL: https://ceur-ws.org/Vol-3287/paper13.pdf.

O-Dang at HODI and HaSpeeDe3: A Knowledge-Enhanced Approach to Homotransphobia and Hate Speech Detection in Italian

Chiara Di Bonaventura[1], Arianna Muti[2] and Marco Antonio Stranisci[3]

[1]King's College London, United Kingdom
[2]University of Bologna, Italy
[3]University of Turin, Italy

Abstract
This paper describes our methods implemented during the EVALITA 2023 campaign for homotransphobia (HODI task) and hate speech detection (HaSpeeDe3 task) in Italian. We present three knowledge-enhanced approaches, namely via triple verbalisation, via prompting and via a majority vote, and we compare them to the AlBERTo baseline. These systems leverage the knowledge graph O-Dang, which contains information about named entities in Italian dangerous speech. Our knowledge-enhanced systems outperformed all the competition's baselines. Our best submissions achieved the macro-F1 score of 0.912 for HaSpeeDe3 and 0.795 for HODI, reaching the 1st and 3rd place, respectively. These results were achieved by using our baseline for HODI, and a majority voting approach for HaSpeeDe3.[1]

Keywords
hate speech, knowledge graph, entity linking, data augmentation, prompting

***Warning**: This paper contains examples of potentially offensive content.*[1]

1. Introduction

Technological progress and increasing online communication have made it necessary to create automatic tools for the detection of online abusive language to protect users. Indeed, online abusive language not only has increased over the past years, but also has effects that usually expand beyond the online context[2]. Usually, most of the targeted victims belong to minority groups because of their gender identity, sexual orientation, political and religious affiliation, *inter alia*. Therefore, their protection is of the utmost importance. Gender- and sexual-based violence manifests in social networks every time the abusive language harms LGBTQIA+ individuals directly or indirectly with homotransphobic discourse [2]. Political- and religious-based hate speech instead discriminates people based on their beliefs, affiliations, or ideologies.

Within the NLP community, many recent works propose solutions to automatically classify misogynous and sexist content [3, 4, 5], homotransphobic content [6, 7], and religious hate speech [8].

Evaluation campaigns have sped up the development of innovative approaches for shared tasks. One example is the Homotransphobia Detection in Italian (HODI) shared task [9], which was introduced during EVALITA 2023 [10]. The task comprises two subtasks. The first subtask, Subtask A - Homotransphobia detection, focuses on automatically identifying whether Italian online posts contain homotransphobic content or not. Subtask B - Explainability, aims at extracting the rationales of the classification model trained for Subtask A.

Another example is the shared task HaSpeeDe3 [11], presented during EVALITA 2023 to boost research on political and religious hate speech in Italian tweets. The task counts two subtasks. The goal of the first subtask, Subtask A - Political Hate Speech Detection, is to determine whether the message contains political hate speech or not. The problem is further divided in two distinct approaches, namely textual and contextual. In the former, participants can only use the provided textual content of the tweets, whereas in the latter, they can employ additional contextual information (e.g., metadata of the tweet and author, friends, etc.). Subtask B - Cross-domain Hate Speech Detection, instead, proposes a binary hate speech detection on test data belonging to different hate domains. Precisely, two settings are evaluated: XPolitical-Hate, where the participants can use external data from any kind of other domains, and XReligiousHate, where submissions are tested on tweets from the religious hate domain.

In this paper, we describe our proposed approach to ad-

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
chiara.di bonaventura@kcl.ac.uk (C. D. Bonaventura); iarianna.muti2@unibo.it (A. Muti); marcoantonio.stranisci@unito.it (M. A. Stranisci)
© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

[1]Profanities have been obfuscated with PrOf (https://github.com/dnozza/profanity-obfuscation) [1]

[2]https://www.ohchr.org/en/stories/2021/03/report-online-hate-increasing-against-minorities-says-expert

CEUR-WS.org/Vol-3473/paper29.pdf

dress the HODI and HaSpeeDe3 shared tasks. We propose a knowledge-enhanced approach on top of the AlBERTo baseline that leverages external knowledge (System 1), internal knowledge (System 2) or both (System 3). Our results suggest that knowledge-enhancement can improve abusive language detection depending on the hate domain under study.

The rest of the paper is structured as follows. Section 2 describes the training datasets provided by the tasks' organisers whereas Section 3 describes our proposed systems. Section 4 summarises the experiments performed and discusses the results. Section 5 includes related work in the field. Section 6 draws some conclusions and discusses further possible research lines.

2. Data

To address the two tasks, datasets were provided by the tasks organisers. For HODI, we focus on Subtask A. 5,000 tweets were provided, manually labelled according to two classes, homotransphobic and not. Data are slightly skewed towards the negative class. For HaSpeeDe3, we focus on both Subtask A and B, where A aims to classify tweets between hateful and not, where hate is addressed towards politicians, while in Subtask B hate is addressed towards religious communities, in a cross-domain setting. PolicyCorpusXL (for Task A) contains 7,000 tweets about political debates, while ReligiousHate (for Task B) is composed by 3,000 tweets about the three main monotheistic religions, namely Christianity, Islam and Judaism.

3. Description of our Systems

In this work, we adopt a knowledge-enhanced approach to address the following subtasks: Subtask A for the HODI shared task, and Subtask A - textual and Subtask B - XReligiousHate for the HaSpeeDe3 shared task. However, the systems submitted for the Subtask A - textual satisfy the constraints for Subtask A - contextual and Subtask B - XPoliticalHate too.

Our intuition is to leverage knowledge about named entities in the training data and their association to online abusive language, which can provide auxiliary information to solve the tasks. Abusive language detection systems often fail to capture different nuances of abusive language because of the lack of contextual information [12] and the target-oriented nature of hate speech [13]. We propose knowledge-injection of relevant information into the system to help address this challenge. Firstly, we link training instances to the named entities associated with them. Then, we collect relevant information about named entities from the O-Dang knowledge graph and the Davinci OpenAI model, which are then injected into the AlBERTo baseline [14], a version of BERT for the Italian language trained on Twitter posts which includes emojis, links, hashtags, and mentions. AlBERTo was trained on 200M tweets randomly sampled from the TWITA corpus [15].

The following paragraphs describe the entity linking step, explain the knowledge-enhancement methods we implemented, and provide insights on its application to HODI and HaSpeeDe3 shared tasks.

3.1. Entity Linking

Our knowledge augmentation strategy relies on an entity linking pipeline based on a Knowledge Graph (KG) modelled on the Ontology of Dangerous Speech (O-Dang) [16]. The KG is a snapshot of a Wikidata [17] dump[3] where only entities of the type 'person' were retained, together with a set of 31 properties conveying relevant information (eg: 'member of political party' (P102), 'place of birth' (P19), spouse (P26)). As a result, we obtained 9,552,706 entities and 69,521,846 triples related to them.

After building the knowledge base, we implemented a pipeline for Entity Linking organized in three steps:

1. we identified all the entities of the type PERSON in the training sets of HaSpeeDe and HODI with Spacy[4];
2. we found all potential candidates of each detected entity, by using the Wikipedia APIs[5];
3. we generated three scores for each pair of the type <entity, candidate>: string similarity based on the Ratcliff/Obershelp pattern recognition [18], cosine similarity based on AlBERTo embeddings [14], the ranking of candidates returned from Wikipedia APIs. We retrieved 10 candidates and we only kept the most relevant result.

An example of such a pipleline is the following: Spacy identified 'Zorzi' as a named entity of the type person in the tweet: '@user_abcdefghij **Zorzi**, poveretto parla come una ch*cca isterica *@user_abcdefghij **Zorzi**, poor guy, he talks like a f*ggot*'. We queried Wikipedia APIs inputting the string 'Zorzi', obtaining the candidate 'Tommaso Zorzi', which was the first ranked of the list. We computed the Ratcliff/Obershelp pattern recognition to obtain a similarity score between 'Zorzi' and 'Tommaso Zorzi', which was 0.55 and we did the same with their embeddings (0.9). We averaged these scores to obtain a general score of 0.818. After a manual review of the scores, we decided to keep only linked entities with a score equal to or above 0.8.

[3] https://academictorrents.com/download/229cfeb2331ad43d4706efd435f6d78f40a3c438.torrent

[4] https://spacy.io/

[5] https://www.mediawiki.org/wiki/API:Search

The resulting number of entities identified in the two datasets and linked to our KG are 388 from HODI and 556 from HaSpeeDe3.

3.2. Knowledge-enhancement

We compare the AlBERTo baseline to the following three knowledge-enhanced systems:

System 1: enhanced-AlBERTo with triple verbalisation. For each named entity in the input data, we create a verbalised description using the information retrieved from O-Dang [16] through entity linking. Then, this description is concatenated at the end of the input instance, and passed to AlBERTo for the classification. We generate up to three different verbalised templates per O-Dang property to account for linguistic variation. For instance, the property 'P108' which stands for job location is translated into the following templates: (i) 'lavora presso' *(works at)*, (ii) 'svolge la sua professione presso' *(carries out their profession at)*, or (iii) 'svolge il suo lavoro presso' *(carries out their job at)*. When creating these templates, we ensure gender neutrality to avoid any possible source of bias. Then, for each input naming an O-Dang entity, the system randomly picks one template per property to create the verbalisation of the O-Dang triples whose heads refer to the named entity. As a result, the verbalised descriptions for the same named entity tend to be all linguistically different. Further, we fine-tune Italian BERT[6] on ItaCoLa [19] to check the linguistic acceptability of the descriptions. All descriptions passed the test since they were predicted as linguistically acceptable according to the model.[7] Table 1 shows statistics of the O-Dang triples and their verbalisation: the average number of triples, the max number of triples, and the average length of the verbalised triples of an O-Dang entity.

	HaSpeeDe3	HODI
avg #triples	9.9	9.7
max #triples	24	24
avg length verbalised triples	43	40

Table 1
Statistics of O-Dang triples and their verbalisation

System 2: enhanced-AlBERTo with prompting Davinci model. For the second system, we experiment with prompting large language models to leverage the knowlegde about abusive language they have acquired during pretraining. We try the following prompts, among others:

- **P1:** quanto è probabile che ENTITY scriva un tweet offensivo? Se sì, perchè? *(How likely is ENTITY to write an offensive tweet? If so why?)*
- **P2:** quanto è probabile che ENTITY sia associato ad un tweet offensivo? Se sì, perchè? *(How likely is ENTITY to be associated to an offensive tweet? If so why?)*
- **P3:** quanto è probabile che ENTITY sia vittima di un tweet offensivo? Se sì, perchè? *(How likely is ENTITY to be victim of an offensive tweet? If so why?)*

Among them, only prompt **P3** returned relevant results. Indeed, direct questions such as **P1** and **P2** triggered the language model into answering with a non-response — i.e., 'Come modello di intelligenza artificiale, non ho accesso a informazioni specifiche sulle associazioni di ENTITY con tweet offensivi' *(As an artificial intelligence model, I do not have access to specific information about ENTITY's associations with offensive tweets)*. On the other hand, the keyword 'victim' does not trigger the model, which is thus able to return reasonable answers given its pretraining knowledge. Further, we compared popular large language models in order to choose the best one. Table 2 and Table 3 show the difference in returned answers between Davinci and Curie OpenAI models, respectively. Davinci performs very well, whereas Curie – which according to the OpenAI website[8] performs almost equally well – performs poorly with respect to knowledge about offensive content.

Similar to System 1, Davinci's knowledge is injected into AlBERTo by concatenating the returned answer to prompt P3 at the end of the input instance, where ENTITY refers to the named entity in the instance. If the instance names multiple entities, we concatenate all their descriptions.

System 3: Majority vote. The third system takes a majority vote between the AlBERTo baseline, the enhanced-AlBERTo with triple verbalisation and the enhanced-AlBERTo with prompting Davinci. The intuition behind is to leverage both external knowledge provided by O-Dang triples and internal knowledge provided by Davinci model.

3.3. Approach to HODI Shared Task

Subtask A. In this task, we compare AlBERTo baseline to enhanced-AlBERTo with triple verbalisation (System 1) and enhanced-AlBERTo with prompting Davinci model (System 2).

[6] https://huggingface.co/dbmdz/bert-base-italian-xxl-cased

[7] The fine-tuned BERT we used for the inference on the verbalised descriptions achieved a total MCC score of +0.57 on the ItaCoLa dataset.

[8] https://platform.openai.com/docs/models/gpt-3

	Davinci	Curie
OpenAI descriptions	Most capable GPT-3 model. Can do any task the other models can do, often with higher quality.	Very capable, but faster and lower cost than Davinci.
Prompt P3 with ENTITY = Matteo Salvini	Il leader della Lega Nord, Matteo Salvini, è probabilmente vittima di commenti online offensivi a causa della sua politica anti-immigrazione. Salvini ha preso una posizione molto dura contro l'immigrazione clandestina e ha fatto campagne per limitare l'ingresso di immigrati in Italia. La sua politica ha attirato molte critiche da parte di coloro che sostengono che l'immigrazione è un diritto umano fondamentale. Inoltre, Salvini è spesso criticato *The leader of Lega Nord, Matteo Salvini, is probably the victim of offensive online comments because of his anti-immigration policy. Salvini took a very tough stance against illegal immigration and campaigned to limit the entry of immigrants into Italy. His policy has attracted much criticism from those who argue that immigration is a fundamental human right. In addition, Salvini is often criticized*	Se è probabile che un commento online offensivo possa costare una condanna penale, perché non si cerca di intervenire? Perché non si fa niente? Perché non si fa niente? Perché non si fa niente? Perché non si fa niente? *If an offensive online comment is likely to cost a criminal conviction, why is there no attempt to intervene? Why is nothing done? Why is nothing done? Why is nothing done? Why is nothing done?*

Table 2
Comparison between Davinci and Curie OpenAI models on how they reply to the question 'How likely is Matteo Salvini to be victim of an offensive tweet? If so why?' (P3)

	Davinci	Curie
Prompt P3 with ENTITY = Matteo Renzi	Nel corso degli ultimi anni, Matteo Renzi è stato vittima di diversi commenti online offensivi. In particolare, nel 2016, è stato preso di mira da alcuni utenti che hanno criticato il suo modo di gestire la crisi economica in Italia. *Over the past few years, Matteo Renzi has been the victim of several offensive online comments. In particular, in 2016, he was targeted by some users who criticized his way of managing the economic crisis in Italy.*	Il commento è stato dato da una persona molto influente, che è il presidente del partito, che ha dato una risposta che l'ha definita "inaccettabile". Il partito non è una scuola di formazione, ma una scuola di vita, e quindi la risposta deve essere meno formale, e più educata. Perché questa persona è il presidente del partito? C'è una persona che rappresenta il partito, che è una persona che rappresenta la linea *The comment was given by a very influential person, who is the president of the party, who gave a response that called it "unacceptable". The party is not a training school, but a school of life, and therefore the answer must be less formal, and more educated. Why is this person the party president? There is a person who represents the party, who is a person who represents the political line*

Table 3
Comparison between Davinci and Curie OpenAI models on how they reply to the question 'How likely is Matteo Renzi to be victim of an offensive tweet? If so why?' (P3)

3.4. Approach to HaSpeeDe3 Shared Task

Subtask A - textual. For hate speech detection in the political domain, we submit our System 2 (baseline enhanced with Davinci) and the majority vote (System 3).

Subtask B - XReligiousHate. Similarly, we compare System 2 and System 3. Since subtask B explores cross-

domain hate speech detection by focusing on the religious domain, we further augment the training dataset with CONAN [8], which is a multilingual expert-based hate speech/counter-narrative pairs dataset on Islamophobia. More precisely, we use only Italian hate speech content for augmentation.

4. Experimental Setup and Results

Setting For the baseline, we fine-tune AlBERTo to our downstream tasks. We perform a minimum parameter selection tuning on the validation set (10% of the training set). We selected the highest performing learning rate $\in [1e{-}5, 2e{-}5, 1e{-}2]$; batch size $\in [4, 8, 16, 32]$; epochs in range $[1 - 10]$. The best configuration for both models is: lr = 1e−5, batch size = 16, epochs = 4. In order to tune the network, we used the AdamW optimizer. As for the pre-processing, we used the pretrained AlBERTo tokenizer for text tokenization, and then we encoded the data. We set the maximum length to 256 characters for the baseline, and 512 for instanced enriched with verbalisation or prompting.

For System 2, we use *text-davinci-002* with *temperature=0.5*. To avoid too long answers, particularly for multiple named entities, we set the max number of returned sentences to *n=5* with *max_tokens=150*.

Results Tables 4 and 5 show the results on the development and test set of the HODI and HaSpeeDe3 shared tasks. While for HODI we submitted all models, for HaSpeeDe3 only the models that performed the best in the development set were subsequently submitted. For HaSpeeDe3 - Task B we could not produce results on the development set because the religious data were only released during the testing phase.

Setting	Dev	Test
AlBERTo (run 1)	0.83	0.795
Davinci (run 2)	0.87	0.792
Verbalisation (run 3)	0.82	0.780
Majority Vote (not submitted)	-	0.80
baseline	-	0.669
top 1	-	0.810

Table 4
Experiments and macro-averaged F_1 scores on the development and test set for HODI. We report also the scores of the baseline produced by the organisers and the top-performing team.

In both tasks, our models struggle the most with the identification of the positive class, which is the least represented in the training data. For HODI, enhancing the baseline with the text generated from Davinci helps us gain 0.04 points with respect to the baseline in the development set, but we observe a slight drop in the test set, leading our baseline to be our best submitted system. In both the development and test set, enhancing the baseline with the verbalisation cause a drop of 0.01 point with respect to the baseline. Majority voting does not prove to be effective in this case. For HaSpeeDe3, we obtain the best result by enhancing our baseline with Davinci, gaining 0.01 point. Combining both the verbalisation and Davinci leads to a drop of 0.01 point instead, while adding only the verbalisation does not affect the final score. However, in the test set we obtain the best result with majority vote, gaining two units over the best-performing model, i.e., Davinci.

	Task A		Task B
Setting	dev	test	test
AlBERTo	0.91	0.91	0.51
Davinci (run 2)	0.92	0.89	0.48
Verbalisation	0.91	0.91	0.51
Davinci + Verbalisation	0.90	0.89	0.47
Majority Vote (run 1)	-	0.91	0.52

Table 5
Experiments and macro-averaged F_1 scores on the development and test set for Subtasks A and B of HaSpeeDe3.

Error Analysis We performed an analysis of the classification errors over our submitted systems. First, with the help of NLTK, we retrieved the most frequent words in misclassified instances. In both tasks, enhancing the models with a knowledge-base, either through Davinci or the verbalisation, results in less misclassified instances containing name entities, proving the efficacy of our approach, at the expenses of other instances that do not contain name entities hence are not enhanced, which do not get identified correctly. For what concerns HaSpeeDe3, most false positives comes from sentences showing a negative sentiment towards racist statements, like in the following example: *Se la SeaWatch fosse piena di gattini sareste già partiti con i pedalò per salvarli. Mi disgustate.*[9] As for false negatives, we observed a pattern of misclassification when the target of the statement is implicit, like in the following example: *Quelli che si lamentano della puzza di urina per le strade di Roma sono gli stessi che dicono #Fateliscendere o #portiaperti ! Ma secondo loro chi c*zzo è che piscia per strada a tutte le ore, che vagano ubriachi o senza meta? Gli alieni?.*[10] In this case, there is an implicit stereotyped racist statement, i.e., those who pee on the streets are all immigrants.

[9] If SeaWatch was full of kittens you would have already left with paddleboats to rescue them. You disgust me.

[10] Those who complain about the stench of urine on the streets of Rome are the same who say #Letthemin or #opentheports ! But according to them who the f*ck is pissing in the streets at all hours, wandering drunk or aimlessly? The aliens?

5. Related Work

Recent work has shown the efficiency of knowledge-enhancement of NLP models in many downstream tasks, such as sentiment classification [20], word sense disambiguation [21], and semantic change detection [22].

Sharifirad et al. [23] is one of the first attempts to leverage external world knowledge for abusive language detection, improving performance on sexist tweet classification. They use ConceptNet [24, 25] and Wikidata [17] to augment the original data by concatenating additional information about concepts and their descriptions. Similarly, Lin [26] uses an entity linking approach to link entities mentioned in tweets to their Wikipedia descriptions in order to leverage world knowledge for hate speech detection. The injection of external world knowledge is a promising avenue for explicit hate speech detection, leading to an improvement of 10% for precision, recall and F1-score. We build upon these works, and leverage both external and internal knowledge about abusive language to explore their impact on different hate domains. To the best of our knowledge, we are the first to apply internal knowledge of large language models through prompt-engineering to the hate speech detection task, while most of the works focus on natural language understanding, question answering or text completion [27].

6. Conclusions

We present a knowledge-enhanced classification solution for identifying homotransphobic and hate speech content in Italian online posts. Our first system uses external knowledge injection via O-Dang triple verbalisation to enhance the AlBERTo model, whereas our second system exploits Davinci's internal knowledge about abusive language to enhance AlBERTo model. Lastly, a majority vote among AlBERTo baseline, System 1 and System 2 is adopted to improve classification particularly on uncertain predictions. We evaluate our approach in the Homotransphobia Detection in Italian (HODI) and Hate Speech Detection (HaSpeeDe3) Shared Tasks of the EVALITA 2023 campaign. Our results show that knowledge-enhancement can improve the classification, especially of sentences containing name entities in the political hate domain. The resulting approach can be expanded on multiple knowledge sources, knowledge injection methods and tasks.

Acknowledgments

The work of Chiara Di Bonaventura was supported by UK Research and Innovation [grant number EP/S023356/1], in the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence (www.safeandtrustedai.org). CDB would like to thank her supervisors, Albert Meroño-Peñuela and Barbara McGillivray, for their helpful comments and mentorship.

References

[1] D. Nozza, D. Hovy, The state of profanity obfuscation in natural language processing scientific publications, in: Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, 2023.

[2] B. R. Chakravarthi, R. Priyadharshini, R. Ponnusamy, P. K. Kumaresan, K. Sampath, D. Thenmozhi, S. Thangasamy, R. Nallathambi, J. P. McCrae, Dataset for identification of homophobia and transophobia in multilingual youtube comments, arXiv preprint arXiv:2109.00227 (2021).

[3] E. Fersini, D. Nozza, G. Boifava, Profiling Italian misogynist: An empirical study, in: Proceedings of the Workshop on Resources and Techniques for User and Author Profiling in Abusive Language, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 9–13. URL: https://aclanthology.org/2020.restup-1.3.

[4] E. W. Pamungkas, V. Basile, V. Patti, Misogyny detection in twitter: a multilingual and cross-domain study, Information Processing & Management 57 (2020) 102360.

[5] H. R. Kirk, W. Yin, B. Vidgen, P. Röttger, Semeval-2023 task 10: Explainable detection of online sexism, in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, 2023. URL: http://arxiv.org/abs/2303.04222. doi:10.48550/arXiv.2303.04222.

[6] D. Locatelli, G. Damo, D. Nozza, A cross-lingual study of homotransphobia on Twitter, in: Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP), Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 16–24. URL: https://aclanthology.org/2023.c3nlp-1.3.

[7] B. R. Chakravarthi, R. Priyadharshini, T. Durairaj, J. P. McCrae, P. Buitaleer, P. K. Kumaresan, R. Ponnusamy, Findings of the shared task on Homophobia Transphobia Detection in Social Media Comments, in: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, Association for Computational Linguistics, 2022.

[8] Y.-L. Chung, E. Kuzmenko, S. S. Tekiroglu, M. Guerini, CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses

to fight online hate speech, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 2819–2829. URL: https://aclanthology.org/P19-1271. doi:10.18653/v1/P19-1271.

[9] D. Nozza, A. T. Cignarella, G. Damo, T. Caselli, V. Patti, HODI at EVALITA 2023: Overview of the Homotransphobia Detection in Italian Task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[10] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[11] M. Lai, F. Celli, A. Ramponi, S. Tonelli, C. Bosco, V. Patti, Haspeede3 at evalita 2023: Overview of the political and religious hate speech detection task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[12] B. Vidgen, D. Nguyen, H. Margetts, P. Rossini, R. Tromble, Introducing CAD: the contextual abuse dataset, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 2289–2303. URL: https://aclanthology.org/2021.naacl-main.182. doi:10.18653/v1/2021.naacl-main.182.

[13] D. Nozza, Exposing the limits of zero-shot cross-lingual hate speech detection, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics, Online, 2021, pp. 907–914. URL: https://aclanthology.org/2021.acl-short.114. doi:10.18653/v1/2021.acl-short.114.

[14] M. Polignano, P. Basile, M. de Gemmis, G. Semeraro, V. Basile, AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets, in: Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019), volume 2481, CEUR, 2019. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85074851349&partnerID=40&md5=7abed946e06f76b3825ae5e294ffac14.

[15] V. Basile, M. Lai, M. Sanguinetti, Long-term social media data collection at the university of turin, in: Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, 2018.

[16] M. A. Stranisci, S. Frenda, M. Lai, O. Araque, A. T. Cignarella, V. Basile, C. Bosco, V. Patti, O-dang! the ontology of dangerous speech messages, in: Proceedings of the 2nd Workshop on Sentiment Analysis and Linguistic Linked Data, European Language Resources Association, Marseille, France, 2022, pp. 2–8. URL: https://aclanthology.org/2022.salld-1.2.

[17] D. Vrandečić, M. Krötzsch, Wikidata: A free collaborative knowledgebase, Commun. ACM 57 (2014) 78–85. URL: https://doi.org/10.1145/2629489. doi:10.1145/2629489.

[18] J. Ratcliff, D. Metzener, Ratcliff-obershelp pattern recognition, Dictionary of Algorithms and Data Structures (1998).

[19] D. Trotta, R. Guarasci, E. Leonardelli, S. Tonelli, Monolingual and cross-lingual acceptability judgments with the Italian CoLA corpus, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 2929–2940. URL: https://aclanthology.org/2021.findings-emnlp.250. doi:10.18653/v1/2021.findings-emnlp.250.

[20] P. Ke, H. Ji, S. Liu, X. Zhu, M. Huang, SentiLARE: Sentiment-aware language representation learning with linguistic knowledge, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 6975–6988. URL: https://aclanthology.org/2020.emnlp-main.567. doi:10.18653/v1/2020.emnlp-main.567.

[21] J. Zhou, Z. Zhang, H. Zhao, S. Zhang, LIMIT-BERT : Linguistics informed multi-task BERT, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 4450–4461. URL: https://aclanthology.org/2020.findings-emnlp.399. doi:10.18653/v1/2020.findings-emnlp.399.

[22] B. McGillivray, M. Alahapperuma, J. Cook, C. Di Bonaventura, A. Meroño-Peñuela, G. Tyson, S. Wilson, Leveraging time-dependent lexical features for offensive language detection, in: Proceedings of the The First Workshop on Ever Evolving NLP (EvoNLP), Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 2022, pp. 39–54. URL: https://aclanthology.org/2022.evonlp-1.7.

[23] S. Sharifirad, B. Jafarpour, S. Matwin, Boosting text classification performance on sexist tweets

by text augmentation and text generation using a combination of knowledge graphs, in: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 107–114. URL: https://aclanthology.org/W18-5114. doi:10.18653/v1/W18-5114.

[24] H. Liu, P. Singh, Conceptnet—a practical commonsense reasoning tool-kit, BT technology journal 22 (2004) 211–226.

[25] R. Speer, J. Chin, C. Havasi, Conceptnet 5.5: An open multilingual graph of general knowledge, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI'17, AAAI Press, 2017, p. 4444–4451.

[26] J. Lin, Leveraging world knowledge in implicit hate speech detection, in: Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI), Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 2022, pp. 31–39. URL: https://aclanthology.org/2022.nlp4pi-1.4.

[27] R. Brate, M.-H. Dang, F. Hoppe, Y. He, A. Meroño-Peñuela, V. Sadashivaiah, Improving language model predictions via prompts enriched with knowledge graphs, in: Workshop on Deep Learning for Knowledge Graphs (DL4KG@ ISWC2022), 2022.

EMit at EVALITA 2023: Overview of the Categorical Emotion Detection in Italian Social Media Task*

Oscar Araque[1,*,†], Simona Frenda[2,3,†], Rachele Sprugnoli[4,†], Debora Nozza[5,†] and Viviana Patti[2,†]

[1]Intelligent Systems Group, Universidad Politécnica de Madrid, Spain
[2]Università degli Studi di Torino, Italy
[3]aequa-tech srl, Turin, Italy
[4]Università di Parma, Italy
[5]Università Bocconi, Italy

Abstract

The Emotions in Italian (EMit) task is the first edition of a shared task on emotion analysis and opinion mining in Italian messages at EVALITA 2023. EMit presents two subtasks: (i) Subtask A, that consists in an emotion detection challenge, and (ii) Subtask B, that introduces a novel problem of target detection of the expressed emotion. Additionally, EMit challenges systems with a thorough in-domain and out-of-domain evaluation, probing the generalization capabilities of the submitted solutions. In general, 4 teams have participated in Subtask A, achieving a macro-averaged f-score of 0.6028 and 0.4977 in the in-domain and out-of-domain sets, respectively. In Subtask B a team has participated, obtaining 0.6459 in the in-domain set and 0.3223 in the out-of-domain set as macro-averaged f-scores. The obtained results indicate that further work needs to be done to solve the task, opening new avenues of research.

Keywords

Emotion detection, Emotion target detection, User-generated contents, Sentiment Analysis

1. Introduction and Motivations

The detection of emotions in texts has a long history in international evaluation campaigns but has never been addressed in EVALITA where the only shared task to deal with emotions was about emotional speech recognition systems [1]. The *Affective Text* shared task at SemEval 2007 was the first one to propose the classification of newspaper headlines according to 6 emotions: anger, disgust, fear, joy, sadness, surprise [2]. Then, starting from 2017, this type of evaluation has become very frequent with a particular attention to the processing of tweets and dialogues. For example, *Affect in Tweets* at SemEval 2018 [3] included a subtask about the multilabel detection of 11 emotions in tweets written in English, Arabic, and Spanish whereas the *Emotion Detection* task at TASS 2020 [4] and EmoEvalEs at IberLEF 2021 [5] were only on Spanish tweets [1]. Instead, *EmoContext* at SemEval 2019 [6] and *EmotionX* at the SocialNLP workshop in 2018 [7] and 2019 focused on the emotion classification of dialogues in English. Last year, the *Emotion Classification* shared task at WASSA 2022 dealt with a different genre of text proposing the classification of emotions in essays written in reaction to news articles [8].

In this context, the **EMit (Emotions in Italian)** task[2] aims at providing the first evaluation framework for emotion detection in Italian texts at EVALITA [9], offering novel annotated data available to the community that will foster future research. EMit tackles a comprehensive emotion model that is complemented with additional annotations regarding the scope of opinions.

2. Task Description

EMit is organized according to two subtasks, thus offering participants different perspectives on opinion analysis:

- **Subtask A**
  Emotion Detection (Main Task) The main proposed subtask is the detection of *emotions* in social media messages about TV shows and series

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
*Corresponding author.
o.araque@upm.es (O. Araque); simona.frenda@unito.it (S. Frenda); rachele.sprugnoli@unipr.it (R. Sprugnoli); debora.nozza@unibocconi.it (D. Nozza); viviana.patti@unito.it (V. Patti)
https://gsi.upm.es/oaraque/ (O. Araque); http://www.di.unito.it/~frenda/ (S. Frenda); https://personale.unipr.it/it/ugovdocenti/person/236480 (R. Sprugnoli); https://deboranozza.com/ (D. Nozza); https://www.unito.it/persone/vpatti (V. Patti)
0000-0003-3224-0001 (O. Araque); 0000-0002-6215-3374 (S. Frenda); 0000-0001-6861-5595 (R. Sprugnoli); 0000-0002-7998-2267 (D. Nozza); 0000-0001-5991-370X (V. Patti)
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

[1]https://competitions.codalab.org/competitions/28682
[2]Task website: http://www.di.unito.it/~tutreeb/emit23/index.html, task repository: https://github.com/oaraque/emit

CEUR-WS.org/Vol-3473/paper1.pdf

emitted by RAI (Radiotelevisione italiana, the national public broadcasting company of Italy), music videos and advertisements.
Given a message, the system decides the emotions expressed in the message or the absence of emotions.

- **Subtask B**
  Target Detection The second subtask is about the detection of the target addressed by the author of the message: the *topic* or the *direction*. In each text, it is indicated whether this refers to what the broadcast is about (the *topic*) or whether it refers to something that is under control of the broadcast itself (*direction*). When the target of the post is the *topic*, this means that the text addresses topics such as events, issues discussed in the TV episode/music video/advertisements, or invited guests of a TV show. On the other hand, the target encoded as *direction* implies that the message describes the specific directors of the shows/series, the showman/artists, fixed guests in the TV shows, reporters, or the show/series/-music video/advertisements as such.
  Given a message, the system decides if the target of the message is related to topic, direction, both or none of the two.

Both subtasks are designed as **multilabel problems of classification**. In this way, participating systems are required to provide as output the *id* of the message and all the predicted labels contained in it. It is worth mentioning that in Subtask A, the message may be classified as neutral, or expressing one or more emotions. Thus, the provided labels are: `neutral` when the message does not express any emotion, the 8 main emotions defined by Plutchik in [10] (`anger`, `anticipation`, `disgust`, `fear`, `joy`, `sadness`, `surprise`, `trust`), and the additional label `love` that is one of the primary dyads in the Plutchik's wheel of emotions, being a combination of *joy* and *trust*. Therefore, a total of 10 labels are used for Subtask A. In Subtask B the message can be classified as addressing the topic, the direction, or both or neither, thus the provided labels are: `topic` and `direction`.

Considering the specific attention on the entertainment sector, we designed Subtask B particularly on the events and players involved in such contents and in their creation. Indeed, the combination of the two subtasks allows going beyond the simple detection of emotions, identifying also if the target of the affective comments about TV programs is related to the *topic* or to issues under control of the broadcasting company (the *direction*). Such finer grained information can be of great importance in real application domains, for artists or broadcasters in the evaluation of the contents delivered, when the analysis of emotions in social media is used as a social signal of emotional reactions of Italian television audience. In other words, this would lead to the development of an *Auditel of emotions*.

3. Datasets

In order to evaluate the robustness of the models proposed by participants, in EMit we release two different test sets: (i) the *in-domain* dataset, which including tweets of the same textual genre and subjects of the training set, and (ii) an additional *out-of-domain* set that is composed of social text of different genres and subjects[3]. In this way, we offer to participants a cross-domain evaluation setting for both subtasks A and B. Table 1 summarizes the size and distribution of the datasets used in EMit 2023.

Learning Set	Dataset	Total (approx.)
Subtask A		
Train	In-domain	5,966
Test 1	In-domain	1,000
Test 2	Out-of-domain	1,000
Subtask B		
Train	In-domain	5,966
Test 1	In-domain	1,000
Test 2	Out-of-domain	1,000

Table 1
EMit 2023 datasets and their distribution in subtasks.

Dataset for in-domain evaluation.
This dataset is obtained from Twitter and it is composed of 6,966 tweets that discuss programs by the Italian RAI TV station. Such messages have been grouped in almost 5 set, each set annotated by three different annotators (for a total of 15 annotators) with a multi-layered annotation scheme. As described, the *emotion* layer consists of 10 labels: Plutchik's emotions, `love` and `neutral`. These emotion annotations are used for running Subtask A.

The emotion labels are non-exclusive, thus a certain tweet can be annotated with one or more emotions, or even solely as neutral , as shown in the examples in Table 3. The number of tweets that expresses at least one emotion is the 78% of all tweets, which is a fairly high coverage. Also, the number of tweets that express two or more emotions represents the 19% of all tweets.

On top of this, the dataset is annotated with the innovative layer concerning the *target*, including the `topic` (describing the events of the emission) and `direction` (whether messages are directed to a specific

[3]It is important to note that user data is not disclosed, since all data has been anonymized by removing all personal information such as @usernames and generating new IDs for the texts coming from Twitter.

Label	Train	Test (in-domain)	Test (out-of-domain)
Anger	367	56	122
Anticipation	547	85	45
Disgust	874	165	152
Fear	91	13	2
Joy	650	100	98
Love	633	103	135
Neutral	1,322	210	42
Sadness	545	95	132
Surprise	591	102	89
Trust	1,665	272	584
Direction	1,698	260	890
Topic	3,805	671	236
Neither	978	149	9

Table 2
Number of annotations, detailed by label and dataset fold.

Text	Joy	Ang.	Trust	Neu.	Sad.	Love	Dis.	Sur.	Ant.	Fear	Top.	Dir.
Caspita che meraviglia [Gosh what a wonder] #LAmicaGeniale	1	0	1	0	0	0	0	0	0	0	0	0
Queste persone mi spezzano il cuore [These people break my heart] #amorecriminale	0	0	0	0	1	0	0	0	0	0	1	0
il sabato sera con [Saturday evening with] #albertoangela #viaggiosenzaritorno #leggirazziali - Watching Ulisse	0	0	0	0	0	1	0	0	0	0	0	1
Ma i genitori di questi idioti non li hanno mai mandati a scuola? [But didn't the parents of these idiots ever send them to school?] #pechinoexpress	0	1	0	0	0	0	1	0	0	0	1	0

Table 3
Excerpt of the in-domain dataset. The annotations are the following: Text, Joy, Anger, Trust, Neutral, Sadness, Love, Disgust, Surprise, Anticipation, Fear, Topic, Direction.

entity related to RAI) labels. These annotations offer a novel perspective on the data, allowing participants and, in general, the EVALITA community, to explore the effectiveness of current models to understand such a subtask. In total, 84% of the tweets are annotated with the "topic" or "direction" labels, and 8% of tweets have both labels. These annotations should be used for Subtask B.

Dataset for out-of-domain evaluation.
We provide as a second test set 1,000 out-of-domain instances for both subtasks A and B. This additional dataset is composed of comments to music videos and advertisement posted on YouTube. The selection of the videos followed the same procedure used for the creation of the MultiEmotions-It dataset [11]. Specifically, the videos were manually chosen from the songs of Sanremo Music Festival 2021 and from the most recent advertisements, covering different types of products and services. The annotation was performed manually using the same approach of the in-domain dataset. Examples are given in Table 4. In this way, we propose the use of data from a variety of sources that do not directly address RAI contents, but describe other audiovisual media.

As a summary, Table 2 shows the arrangement of the proposed datasets for subtasks A and B, with the detail for each class.

4. Evaluation

In EMit 2023, participants are allowed to submit up to 2 runs for each subtask, with a mandatory run for the main Subtask A. The first run is required to be a constrained submission. That is, the only annotated data to be used for training and tuning the systems are those distributed by the organizers, with the exception of additional data such as lexicons and word embeddings. On the contrary, the second run of each participant can be unconstrained, thus allowing participants to use additional training data. The performance of the systems is evaluated using the **macro-averaged F1-score**, which aggregates the classification metrics for each of the classes thus, in the official

Text	Joy	Ang.	Trust	Neu.	Sad.	Love	Dis.	Sur.	Ant.	Fear	Top.	Dir.
Vergognatevi! Che schifo [Shame on you! How disgusting]	0	1	1	0	0	0	1	0	0	0	1	0
sento la mancanza delle mie crociere .grazie del video e speriamo presto di partire. [I miss my cruises .thanks for the video and hope to go soon.]	0	0	0	0	1	0	0	0	1	0	0	1
Ma quanto è bello Damiano raga? Canzone spaziale ! [But how beautiful is Damiano raga? Space song !]	0	0	1	0	0	0	0	0	0	0	1	1
Adoro questa canzone complimentissimi [I love this song congratulations]	1	0	1	0	0	0	1	0	0	0	1	1

Table 4
Excerpt of the out-of-domain dataset. The annotations are the following: Text, Joy, Anger, Trust, Neutral, Sadness, Love, Disgust, Surprise, Anticipation, Fear, Topic, Direction.

ranking, participants' runs are ordered according to the mentioned F-score.

As baseline, we provide the results of three basic models. All these models compute different text representations that are fed to a logistic regression classifier. In this way, the baselines' text representations are:

- *Baseline_OHE*: uni and bi-grams encoded with a one-hot schema, with a vocabulary of 5,000 tokens.
- *Baseline_TFIDF*: uni and bi-grams represented with the TF-IDF approach, again using a vocabulary of 5,000 tokens.

Finally, we also consider the results of a simple random baseline *Baseline_random*, that outputs the predictions for all classes following a uniform random distribution.

5. Task Overview: Systems and Results

In this first edition of EMit, very few teams participated in the competition. In particular, we received 1 submission by industry (App2Check) and 3 by academic teams (extremITA, ABCD, and EmotionHunters). Although the few participants, the organized shared task also collected international interest with the ABCD team coming from Vietnam. All 4 participating teams have submitted at least one run for Subtask A, and just one team sent us the predictions on Subtask B.

5.1. Systems

Attending to the various systems employed for the classification of emotions (multilabel) and target (binary), their design is based mainly on the use of Large Language Models (LLMs), confirming the actual tendency and success of transformer-based models. However, they have been included in different architectures.

The most used approach is supervised, with a predominance of fine-tuning actions of LLMs to address the specific task of classification. Moreover, two teams also presented semi-supervised systems based specifically on a few-shot prompting (extremITA and App2Check). Various LLMs are employed. For instance, some teams experimented with the classic BERT-based models for the Italian language (i.e., `bert-base-italian-cased`, `bert-base-italian-xxl-cased`, `bert_uncased_L-12_H-768_A-12_italian_alberto`, `umberto-commoncrawl-cased-v1`), others with already fine-tuned versions of BERT (i.e., `feel-it-italian-emotion`, `polibert_sa`), and the rest exploited some sequence-to-sequence LLMs oriented, in this context, to perform mainly instruction solutions such as ChatGPT (`gpt-3.5-turbo-0301`), `flan-t5-xl`, `mt5-base`, IT5 (`it5-efficient-small-el32`) and LLaMA foundational model (`llama-7b-hf`).

In particular, EmotionHunters [12] performed a battery of experiments with classic BERT models and already fine-tuned versions of LLMs. The final system, selected on the basis of their experiments, is based on the fine-tuning of AlBERTo model and, at the top, the fully connected layer to provide a multilabel classification for each text. Both ABCD [13] and App2Check [14] teams employed an ensemble of predictions of different LLMs based on a soft voting method that considers the confidence score associated with each prediction (ABCD: run 1) and the best top-performing model for each emotion (App2Check: unsubmitted run)[4] looking at the performance in the development set of the two best implemented systems: A2C-mT5-r1 (App2Check, run 1) and A2C-GPT-r2 (App2Check, run 2). A2C-mT5-r1 is based on the fine-tuning of multilingual T5 employing the Simple Transformers library. While A2C-GPT-r2 is built using a few-shot approach with ChatGPT, prompt to simultaneously identify all emotions for each text input. A similar approach is used by extremITA [15], who em-

[4]The unsubmitted run reported very good scores in both in-domain (f1-score of 0.504) and out-of-domain setting (f1-score of 0.518)

Ranking	Team name	run id	Anger	Anticipation	Disgust	Fear	Joy	Love	Neutral	Sadness	Surprise	Trust	Macro-average
1	extremITA	2	**0.5176**	**0.6420**	**0.6278**	**0.5833**	0.6178	0.5190	**0.7035**	**0.6258**	**0.5059**	0.6854	**0.6028**
2	extremITA	1	0.4815	0.5594	0.5731	0.1429	0.5909	0.4503	0.6565	0.5233	0.4198	**0.6884**	0.5086
3	ABCD	1	0.4706	0.5946	0.5524	0.0000	**0.6429**	0.4586	0.6462	0.5963	0.3810	0.6516	0.4994
4	EmotionHunters	1	0.4596	0.5205	0.5842	0.2400	0.4589	0.5000	0.4319	0.5484	0.4601	0.6319	0.4835
5	App2Check	2	0.4048	0.3814	0.5831	0.2642	0.3614	**0.5463**	0.3465	0.5181	0.1250	0.2108	0.3741
6	App2Check	1	0.3529	0.4149	0.3855	0.4000	0.6122	0.4867	0.5340	0.4339	0.3293	0.5741	0.4523
7	*baseline_TFIDF*		*0.2945*	*0.4444*	*0.4680*	*0.3684*	*0.3493*	*0.3314*	*0.5392*	*0.3360*	*0.3486*	*0.5944*	*0.4074*
8	*baseline_OHE*		*0.2178*	*0.4221*	*0.3526*	*0.2593*	*0.2918*	*0.3032*	*0.4564*	*0.3191*	*0.2158*	*0.5243*	*0.3362*
9	*baseline_random*		*0.1039*	*0.1541*	*0.2683*	*0.0304*	*0.1760*	*0.1529*	*0.2941*	*0.1565*	*0.2013*	*0.3426*	*0.1872*

Table 5
In-domain evaluation for Task A in terms of f1-score.

Ranking	Team name	run id	Anger	Anticipation	Disgust	Fear	Joy	Love	Neutral	Sadness	Surprise	Trust	Macro-average
1	extremITA	2	0.4051	0.4923	0.6684	0.0000	**0.4416**	**0.7552**	**0.6355**	0.3049	**0.4138**	0.8603	**0.4977**
2	EmotionHunters	1	0.3671	**0.6053**	0.6364	0.0000	0.3768	0.5907	0.6250	0.3103	0.3692	**0.8632**	0.4744
3	extremITA	1	0.5027	0.3667	0.6219	0.0000	0.3176	0.7273	0.5634	0.2024	0.3350	0.8545	0.4491
4	App2Check	1	0.2710	0.4301	0.4691	0.0000	0.4167	0.6528	0.3448	0.2653	0.3662	0.8064	0.4022
5	App2Check	2	**0.6379**	0.3256	**0.6790**	**0.1818**	0.2545	0.6381	0.2564	**0.3195**	0.1373	0.3001	0.3730
6	*Baseline_OHE*		*0.3972*	*0.4533*	*0.4197*	*0.0000*	*0.1890*	*0.3218*	*0.1974*	*0.3129*	*0.2812*	*0.7468*	*0.3319*
7	*Baseline_TFIDF*		*0.3342*	*0.4412*	*0.4092*	*0.0000*	*0.1786*	*0.3034*	*0.1778*	*0.2649*	*0.1913*	*0.6869*	*0.2987*
8	*Baseline_random*		*0.1984*	*0.0917*	*0.2188*	*0.0081*	*0.1624*	*0.2208*	*0.0841*	*0.2097*	*0.1336*	*0.5483*	*0.1876*

Table 6
Out-of-domain evaluation for Task A in terms of f1-score.

ployed sequence-to-sequence LLMs for Italian to solve instructions related to specific tasks. They developed two systems to solve different shared tasks of EVALITA 2023: extremIT5 (extremITA, run 1) and extremITLLaMA (extremITA, run 2). The former is an Encoder-Decoder model based on IT5, and trained by concatenating the task name and an example as input (i.e., "EMit: Quando ci sarà l'espulsione di Claudia #ilcollegio [url]") and as output the sequence of labels; in contrast, the latter is an instruction-tuned Decoder model built upon the LLaMA foundational models, therefore the structured prompt is an instruction in natural language like "Which emotions are expressed in this text? You can choose among joy, fear, ...". Differently from the previous editions of EVALITA, in the EMit 2023 shared task it is clear that the attention is only on the LLMs' ability to solve tasks and their integration into the systems' architecture, losing the focus on linguistic features that can represent or infer the emotions in the text. Also, the preprocessing of the text is focused on very few steps, regarding mainly the transformation of emojis in textual descriptions, removing mentions, urls and other symbols.

5.2. Results

Tables 5, 6, 7, and 8 report the official results obtained in EMit 2023 for both subtask A and B. The ranking is based on the macro-averaged F1-score, and considers both the team and the run of each submission. The higher scores for each column are marked in bold, while the lower scores are underlined.

Generally, it is interesting to see that even if the classification problem of Subtasks A and B are very different, the best results for each are similar. Concretely, when considering the in-domain test set, the best submission for Subtask A obtained a macro-averaged score of 0.6028, while for Subtask B it is 0.6459. In the case of the out-of-domain evaluation, the best scored obtained by a team is 0.4977 in Subtask A, and 0.4448 in Subtask B. This decrease in the classification performance when comparing in-domain and out-of-domain evaluations was expected giving that training was performed only on the in-domain data. Additionally, it is worth noticing that even if Subtask A contains 10 possible labels and Subtask B has only 2, their best scores are not that different (a difference of 0.0431).

Following, in relation to the overall results achieved by participants, it can be seen that in Subtask A, both in the in-domain and out-of-domain evaluations, the teams' submissions have obtained better results than the baselines. The best baseline in the in-domain evaluation uses TF-IDF uni and bi-grams, while for the out-of-domain evaluation the uni and bi-grams using one-hot encoding achieves the best result. Regarding Subtask B, the only team that has submitted a run for it has obtained a better score than in the in-domain evaluation.

In contrast, when considering the out-of-domain evaluation in Subtask B, we see that the best baseline is the one that randomly predicts the objective labels. This decreases in the classification performance is seen in the runs but also in the learning-based baselines. This may be explained by considering the distribution of the out-of-domain sets in Subtask B (see Table 2). Indeed, we can observe that in the train and in-domain test sets the prevalent label is `Topic` but, conversely, in the out-of-domain test set the `Direction` label is more frequent. Consequently, it is possible to postulate that systems trained with the Subtask B training set would perform fairly well in the in-domain test set, but worse on the out-of-domain data.

Ranking	Team name	run id	Direction	Topic	Macro-average
1	extremITA	2	0.4855	**0.8064**	**0.6459**
2	extremITA	1	0.4963	0.7699	0.6331
3	*Baseline_TFIDF*		***0.5032***	*0.7336*	*0.6184*
4	*Baseline_OHE*		*0.4651*	*0.7212*	*0.5932*
5	*Baseline_random*		*0.3443*	*0.5614*	*0.4528*

Table 7
In-domain evaluation for Task B in terms of f1-score.

Ranking	Team name	run id	Direction	Topic	Macro-average
1	*Baseline_Random*		*0.6452*	***0.3651***	***0.5051***
2	*Baseline_TFIDF*		*0.5559*	*0.3573*	*0.4566*
3	extremITA	1	**0.6831**	0.2066	0.4448
4	*Baseline_OHE*		*0.4884*	*0.3571*	*0.4228*
5	extremITA	2	0.3275	0.3172	0.3223

Table 8
Out-of-domain evaluation for Task B in terms of f1-score.

Finally, the detailed results of the evaluation offer interesting insights into the models' performance. For example, when considering the effect of the number of instances for each class (Table 2, we see that in Subtask A `Fear` is much less frequent in comparison to the other emotions. Hence, this has an effect on the performance of the systems: in the out-of-domain evaluation (Table 6) the majority of the models obtained a null score in the `Fear` category, thus affecting in a negative way the overall averaged score. Similarly, the most common emotions in Subtask A (`Trust` and `Neutral`) are generally better predicted by the participants' systems.

6. Discussion

The presence of both in-domain and out-of-domain data in the EMit task provides a valuable experimentation setting as proved by the different performances in classification between the two evaluations settings. Since these two types of datasets have been obtained from different sources (see Sect. 3), they represent a diverse collection of cases. In this way, we can evaluate the participants' models in relation to their generalization capabilities.

In fact, we observe a general reduction in the classification metrics when comparing the in-domain and out-of-domain test sets. In Subtask A, with the in-domain set, the average macro f-score of all participants' systems is 0.4868. In comparison, the average metric drops to 0.4393 in the case of the out-of-domain dataset. We can see a similar trend when considering Subtask B, even if just one team has participated. The average score in the in-domain evaluation is 0.6395 and, in the out-of-domain case, 0.3935.

While participants have achieved promising results in the detection of emotions and opinion targets, there is still room for improvement. The large number of emotions considered in Subtask A is indeed a challenge for automatic systems, increasing the difficulty of the task. In comparison, Subtask B has fewer categories, but still, the proposed systems and baselines obtain rather low metrics in the task. Also, we have seen how the representation of the different emotions greatly impacts classification performance. These, along with the generalization difficulties in the out-of-domain set, indicate that the challenge proposed in EMit is not solved. Indeed, future works need to address the shortcomings detected and advance in the generation of systems that are more robust to the frequency of categories in the datasets, as well as the inclusion of domain-specific knowledge that may improve overall results.

7. Conclusions

The first edition of EMit (Emotions in Italian) proposes the assessment of emotions on Italian texts by presenting an interesting challenge that revolves around two subtasks. On one hand, the main task (subtask A) presents a comprehensive emotion annotation set using Plutchik's model, with the addition of the *love* emotion. On the other hand, subtask B introduces a novel classification problem, which addresses the target of the opinion expressed in the textual message. To complement this, we also provide out-of-domain test sets to further obtain insights into the behaviour of the participants' systems.

To advance in the study of opinion mining in relation to emotion, and considering both subtasks, EMit establishes a rich annotation schema for considering the effect of this challenge on automated systems. While only one team participated in subtask B, we believe that the additional perspectives brought by the combined study of emotions and their targets will be the subject of further

studies. As an example, an interesting research avenue could study the variation of emotions depending on the target, and how this affects learning systems. Another potential research direction is the inclusion of linguistic knowledge into the commonly used large language models.

Acknowledgments

The work of Oscar Araque has been partially funded by the Spanish Ministry of Science, Innovation, and Universities through project COGNOS (PID2019-105484RB-I00) and "ETSI Telecomunicación" of "Universidad Politécnica de Madrid" through the initiative "Primeros Proyectos" under "AFRICA – Detecting and Analyzing Affective and Moral Factors in Radicalization and ExtremIsm: a MaChine learning Approach". The work of S. Frenda and V. Patti was partially funded by the Multilingual Perspective-Aware NLU Project in partnership with Amazon Alexa. The work of D. Nozza was partially funded by Fondazione Cariplo (grant No. 2020-4288, MONICA).

References

[1] A. Origlia, V. Galatà, Evalita 2014: Emotion recognition task (ERT), in: Proceedings of the Fourth International Workshop EVALITA 2014, Pisa University Press, 2014, pp. 112–115.

[2] C. Strapparava, R. Mihalcea, SemEval-2007 task 14: Affective text, in: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Association for Computational Linguistics, Prague, Czech Republic, 2007, pp. 70–74. URL: https://aclanthology.org/S07-1013.

[3] S. Mohammad, F. Bravo-Marquez, M. Salameh, S. Kiritchenko, SemEval-2018 task 1: Affect in tweets, in: Proceedings of the 12th International Workshop on Semantic Evaluation, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 1–17. URL: https://aclanthology.org/S18-1001. doi:10.18653/v1/S18-1001.

[4] M. G. Vega, M. C. Díaz-Galiano, M. Á. G. Cumbreras, F. M. P. del Arco, A. Montejo-Ráez, S. M. J. Zafra, E. M. Cámara, C. A. Aguilar, M. A. S. Cabezudo, L. Chiruzzo, et al., Overview of TASS 2020: Introducing emotion detection, in: IberLEF@ SEPLN, 2020.

[5] F. M. Plaza-del Arco, S. M. Jiménez Zafra, A. Montejo Ráez, M. D. Molina González, L. A. Ureña López, M. T. Martín Valdivia, Overview of the emoevales task on emotion detection for spanish at iberlef 2021 (2021).

[6] A. Chatterjee, K. N. Narahari, M. Joshi, P. Agrawal, SemEval-2019 task 3: EmoContext contextual emotion detection in text, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 39–48. URL: https://aclanthology.org/S19-2005. doi:10.18653/v1/S19-2005.

[7] C.-C. Hsu, L.-W. Ku, SocialNLP 2018 EmotionX challenge overview: Recognizing emotions in dialogues, in: Proceedings of the sixth international workshop on natural language processing for social media, 2018, pp. 27–31.

[8] V. Barriere, S. Tafreshi, J. Sedoc, S. Alqahtani, WASSA 2022 shared task: Predicting empathy, emotion and personality in reaction to news stories, in: Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 214–227. URL: https://aclanthology.org/2022.wassa-1.20. doi:10.18653/v1/2022.wassa-1.20.

[9] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[10] R. Plutchik, H. Kellerman, Theories of emotion, volume 1, Academic Press, 1980. URL: https://books.google.it/books?id=TV99AAAAMAAJ.

[11] R. Sprugnoli, Multiemotions-it: A new dataset for opinion polarity and emotion analysis for italian, in: 7th Italian Conference on Computational Linguistics, CLiC-it 2020, Accademia University Press, 2020, pp. 402–408.

[12] G. Calò, F. Massafra, B. De Carolis, C. Loglisci, TASK A at EVALITA 2023: Overview of the Emotion Hunters Approach to the Categorical Emotion Detection in Italian Social Media, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[13] D. Nguyen Ba, U. N. Ngoc Phuong, T. Dang Van, Ensemble Approach for Categorial Emotion Detection in Social Media Messages: EMit at EVALITA 2023, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[14] G. Cageggi, E. Di Rosa, A. Uboldi, Large Language Models for Multilabel Emotion Classification: fine-tuned LLM vs plain FLAN and ChatGPT, in: Proceedings of the Eighth Evaluation Campaign of Nat-

ural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[15] C. D. Hromei, D. Croce, V. Basile, R. Basili, ExtremITA at EVALITA2023: Multi-Task Sustainable Scaling to Large Language Models at its Extreme, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

Emotion Hunters at EMit: Categorical Emotion Detection combining BERT and ChatGPT models

Gianluca Calò[1], Francesco Massafra[1], Berardina De Carolis[1] and Corrado Loglisci[1]

[1]Department of Computer Science, University of Bari, Italy

Abstract

Emotion detection in text plays a crucial role in various applications, such as customer feedback analysis, social media monitoring, or for the analysis of the verbal part of human communication. Deep learning techniques have shown promising results in accurately recognizing and classifying emotions in textual data. This paper describes the approach to categorical emotion detection of the Emotion Hunters team. After a preprocessing phase, a model fine-tuned from AlBERTo together with the ChatGPT APIs was used to address the challenge. The results show that on the out-of-domain test set our approach performed better than on the in-domain one thus showing a good generalization capability.

Keywords

Emotion Detection, BERT, ChatGPT

1. Introduction

Emotion detection in texts has gained significant importance in recent years due to the pervasive presence of digital communication platforms and the wealth of user-generated content. The ability of software to understand and analyze emotions expressed in a text has numerous applications across various domains, including marketing, customer service, mental health, and social sciences. Emotion detection involves the use of Natural Language Processing (NLP) techniques and machine learning algorithms to automatically identify and classify emotions conveyed in textual content. By accurately detecting and interpreting emotions, we can gain a deeper understanding of human experiences, opinions, and attitudes.

As far as emotion detection for the Italian language is concerned, it presents distinctive features which range from morphological to lexical viewpoints. It presents a lot of words with particles in two or three units (e.g., verbal groups), which are difficult to label. The words used to express the same idea can have different types of grammatical categories, they can be nouns and verbs. In addition, they can be associated with general semantic categories or specific categories. Italian is a very rich language with words that hold more than one meaning, which may mislead an automatic emotion detector. Moreover, while many linguistic resources and annotated texts have been generated for wide-coverage languages, such as English, Chinese and Arabic, the same cannot be said for other less-resourced Indo-European languages, such as Italian.

The EMit (Emotions in Italian) Subtask A [1] at EVALITA 2023 [2] aims at detecting emotions in social media messages about TV shows, music videos and advertisements. Given a message, the system has to decide which emotions are expressed in the message or if the message is a neutral one. According to the annotation of the dataset, the problem to address is designed as a multilabel classification one. Therefore, the system, given a message, will classify it and return all the possible labels denoting emotions contained in it. In particular, in Subtask A, the message could be classified as neutral, or expressing one or more emotions in the following set of 10 labels: anger, anticipation, disgust, fear, joy, sadness, surprise, trust, neutral [3], and the additional label love. Our team, the EmotionHunters, addressed this challenge with a two-steps model. After a pre-processing phase, we fine-tuned a model based on *AlBERTo* [4] and test it on the validation set. The performance on the validation set exceeded the baseline of about 16% reaching an accuracy calculated with the weighted F1-score of 0.56. However, when we run the model on the provided in-domain test-set, we noticed that in some cases the model did not make predictions and that there was a high number of neutral predictions on the total of the results. Then, since the beginning of the call for this challenge, ChatGPT became very popular, for these two cases, we integrated the ChatGPT APIs 3.5 [5] and this increased the prediction of the model of 1%. The proposed model has been tested on the two test sets proposed by the challenge: in-domain, including tweets of the same textual genre and subjects of the training set, and another one, out-of-domain, including social data of different genres and subjects. Our approach showed to have a better performance on the out-of-domain test set showing that it is able to generalize with respect

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT

g.calo26@studenti.uniba.it (G. Calò);
f.massafra7@studenti.uniba.it (F. Massafra);
berardina.decarolis@uniba.it (B. D. Carolis);
corrado.loglisci@uniba.it (C. Loglisci)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper2.pdf

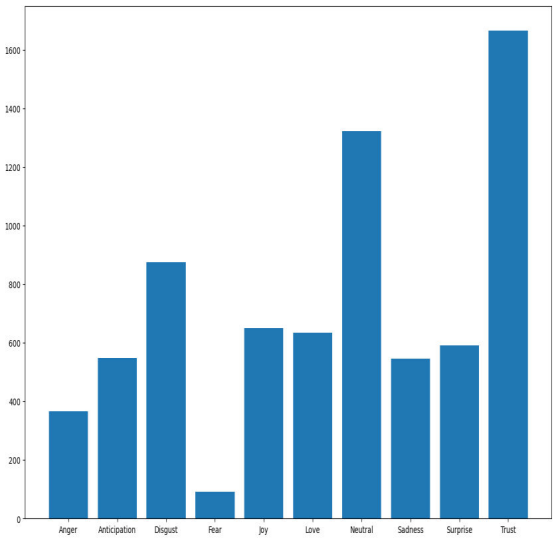


Figure 1: Distribution of examples for each emotion.

to the topic. This is for us an interesting result because we want to apply the model in different domains like the one of conversational experiences with intelligent agents. We did not have the time within the challenge deadline to train a model based only of LLMs like ChatGPT and this is part of our future work.

2. Description of the system

In the following, we first describe the pipeline of text pre-processing and then provide details on the classification algorithms used and configured for the task at hand.

2.1. Analysis of the Dataset

The provided training dataset consists of a collection of 5966 labeled tweets, each identified with one or more labels related to the predicted emotions (among the 10 emotions mentioned above) (see Figure 2). The class distribution is not homogeneous, with the trust and neutral classes being predominant, while the rear class is the least frequent (see Figure 1).

To augment the number of sentences of the fear class, we integrated the training dataset with sentences taken by the dataset MultiEmotions-It [6], moreover, using the affective dictionary proposed in [7], we changed affective terms with synonyms in this way the size of the fear class was upsampled to 400 sentences.

2.2. Pre-processing

A pipeline of preliminary operations is performed in order to first clean and standardize the input messages and then prepare them for a format suitable to the selected learning algorithms [8]. More precisely, we remove the symbols used in social media communication (emoticons, url, links, mentions) without discarding the relative contents but we assign them to semantic categories denoted as tags. For instance, the mentions "@NickName" are converted into the tokens with the uniform and generic tag $< user >$. Each emoticon is converted into the textual description of its meaning taken from a predefined collection of emoticon-description pairs we made for the purpose of this work. For instance, the emoticon with grinning face with big eyes would be converted into the description *<faccina con un gran sorriso e occhi spalancati>* (in Italian). Also, words with hashtags are split into the single tokens, which are then reported with the open and close tags $< hashtag >$ and $< /hashtag >$. For instance, the hashtag "#Sanremo2020" would be converted into the tagged sequence $< hashtag >$ *Sanremo 2020* $< /hashtag >$ composed of two single tokens. The rationale behind these operations is to make tokens and symbols expressing emotions explicit and jointly to augment the features describing the original text. So, the learning process can work on multiple sources of information and better capture the emotive content.

Next, we perform a conversion operation to represent the tweets pre-processed in the input format for the selected learning algorithms, that is, BERT models and variants (as explained in the following). All the tokens

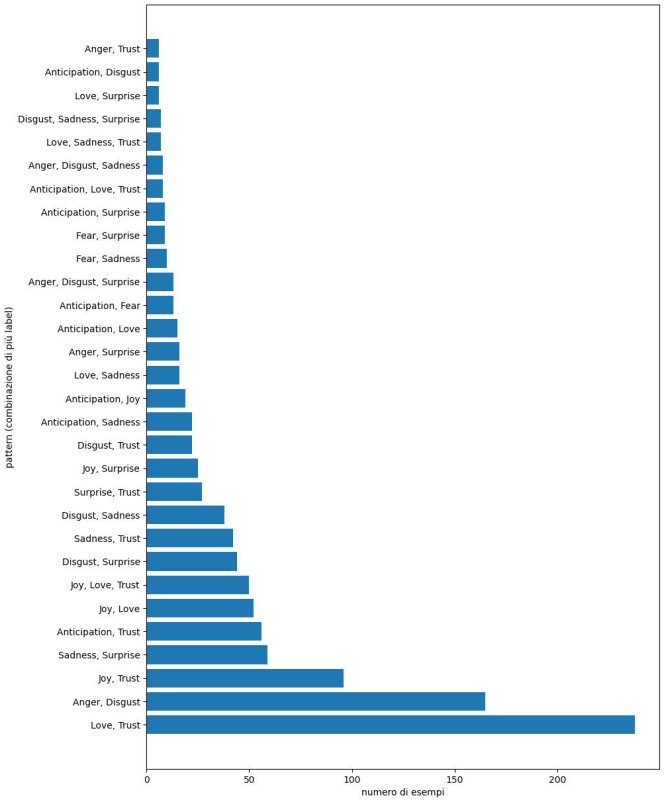


Figure 2: Pattern of labels.

produced for the pre-processed tweets were indexed and used to create a dictionary for the input vectors to the learning process. Considering the typical length of the tweet, usually very less the maximum number of admissible digits, we chose an input length of 128 (elements of the vectors) and prepared an attention mask to decrease the importance of the elements inserted into the input data for padding. Each input vector is in binary code and each element represents the presence/absence of the corresponding indexed token.

2.3. Selected Models

The experimentation is based on BERT models (and variants), which have achieved state-of-the-art results in text classification tasks. BERT utilizes special tokens [CLS] and [SEP] to indicate the beginning of the input sequence and the separation between sentences.

In this specific case, the contextualized embedding associated with the [CLS] token is used as the embedding for the entire sentence. Thanks to the multi-head attention mechanism, it can capture the semantics of the entire sentence effectively.

Several instances of pre-trained BERT models that contain at least some Italian text in their training corpus were considered. For each model, a fully connected layer was added to perform multi-label classification and fine-tune the models on the specific dataset. The models considered are:

- dbmdz/bert-base-italian-xxl-cased [9]: The MDZ Digital Library team released "Italian BERT cased XXL," a BERT version pre-trained on two corpora. The first corpus consists of texts obtained from a Wikipedia dump and various texts from the OPUS collection (http://opus.nlpl.eu/), with a total corpus size of approximately 13 GB and over 2 billion tokens. The second corpus is the Italian part of the OSCAR corpus (https://traces1.inria.fr/oscar/), with a final corpus size of approximately 81 GB and over 13 billion tokens. The "cased" version was chosen as it aligns better with the chosen pre-processing method, as previously done in [8].
- AlBERTo-Base, Italian Twitter lower-cased [4]: A BERT model trained on a corpus of 200 million Italian tweets.
- UmBERTo-Commoncrawl-Cased [10]: A RoBERTa model (a variant of BERT) trained on an Italian sub-

corpus of OSCAR as the training set. It uses a ten-fold version of the Italian corpus, which consists of 70 GB of raw text data, 210 million sentences, and 11 billion words. The sentences were filtered, shuffled at the line level, and utilized for NLP research.

- MilaNLProc/feel-it-italian-emotion [11]: An adapted version of UmBERTo for classification on the Feel-IT dataset.
- bert-base-multilingual-uncased [12]: A multilingual uncased BERT model.

2.4. Training the Models

The challenge provides two baselines and the corresponding code to reproduce their execution. The first baseline uses count vectors to represent the documents based on token frequency, while the second baseline uses TF-IDF vectors. Both implementations limit the vector dimensions to 5000. In the emotion recognition task, the TF-IDF baseline performs better, achieving an F1 score of 41.48%.

For both the baselines and the experiments conducted in this work, a seed was fixed to ensure reproducibility.

Taking into consideration the work presented in GoEmotions [13], we decided to freeze the layers of the pre-trained BERT model and train only the additional classification layers.

Various preliminary experiments were conducted by manually modifying the model's hyperparameters, such as the number of epochs, batch size, learning rate, etc., to understand which was the better approach for this task. To improve the results, the following decisions were made: The Optuna library was adopted to systematically test different combinations of hyperparameters. The MLFlow library was used to track intermediate (F1) and final results (F1 and metrics for each class).

The hyperparameter search space was defined as follows:

- Learning rate: between 2e-05 and 5e-05
- Epsilon (AdamW optimizer): between 1e-8 and 1e-6
- Hidden dropout probability: between 0.1 and 0.3
- Patience for early stopping: a discrete interval between 1 and 5
- Batch size: 16, 32, and 64

The numerous trials conducted using Optuna allowed us to observe that the models pre-trained consistently on an Italian text corpus performed better than those pre-trained on a multi-lingual corpus, including Italian.

In general, it is observed that training that exceed the fourth epoch often result in a degradation of performance in terms of F1 score, and the ideal batch size was found to be 16. On average, the performance of all models benefits from the preprocessing step, as this strategy likely retains more informative content and includes typical social media expressions in standard Italian [14]. A significant contribution can be attributed to the translation of emojis into their Italian descriptions, which can be discriminative in emotion recognition. As a resource we used the one in [15] whose *CLDR Short Name* was translated into Italian.

The best trials were achieved with the BERT AlBERTo model, with learning rates ranging from 2e-05 to 3e-05, a patience value of 3 allowing for training for 4 epochs, a batch size of 16, and employing the "transform" preprocessing strategy. The best trial, in particular, was achieved with the following hyperparameters:

- Learning rate: 2.6263880993374053e-05
- Epsilon: 1.7454440554963499e-7
- Hidden dropout probability: 0.2
- Patience for early stopping: 3
- Batch size:

AlBERTo and dbmdz had a similar performance, however, we selected AlBERTo with an average F1 score of 0.562 also because it had a better performance in classifying fear, which was one of the most problematic due to the limited number of examples (see Figure 3).

2.5. Prediction

During the prediction phase on the test dataset made using the model fine-tuned from AlBERTo, we noticed a high number of neutral examples, and in some cases, the model was unable to determine the emotion, leaving the result field empty. We noticed this problem only on the test set and not on the validation set, therefore, to address this issue, in addition to using our trained model, we integrated the ChatGPT APIs to obtain additional results to fill in the case of neutral or undetermined sentences. Then, each sentence in the test set is pre-processed and given as input into the fine-tuned model, with a threshold set at 0.5. At the end of the prediction phase, every sentence that didn't receive any label or was classified as neutral is passed to a Python program that utilizes the ChatGPT APIs 3.5. Below, we show the prompt used and some of the examples provided to ChatGPT. Due to the limitations of the free APIs, the number of tokens and the amount of input examples are limited.

```
prompt = "Your are an emotion recognition tool for
↪  tweets and your task is to " \
  "analyze them and give a single emotion or a list of
  ↪  emotions, separated by comma that you might think
  ↪  are expressed in the current tweet and you should
  ↪  use only the emotions from " \
  "this list ['anger', 'anticipation', 'disgust', 'fear',
  ↪  'joy', 'love', 'neutral', 'sadness', 'surprise',
  ↪  'trust']"

  examples = """ \
  Here some examples:
  Input "Io ancora non ho capito se la voce mentre
  ↪  cantano sia modificata o meno
  ↪  #IlCantanteMascherato" Output:neutral
  Input "RT @user: Tartaruga is the new zoccola enorme
  ↪  #chilhavisto" Output:disgust
```

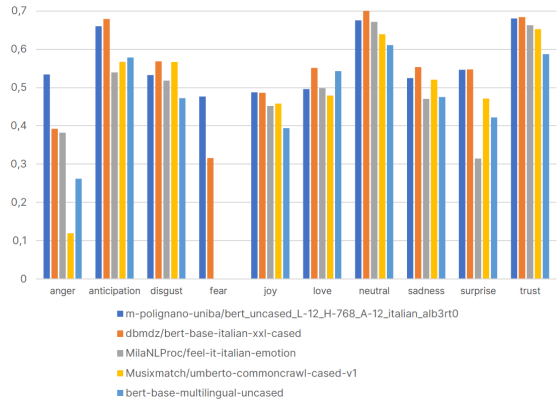


Figure 3: Comparison of the models performance.

Table 1
F1-score evaluation for in-domain testing

team	run id	anger	anticipation	disgust	fear	joy	love	neutral	sadness	surprise	trust	macro-avg
extremITA	2	0,5176	0,6420	0,6278	0,5833	0,6178	0,5190	0,7035	0,6258	0,5059	0,6854	0,6028
extremITA	1	0,4815	0,5594	0,5731	0,1429	0,5909	0,4503	0,6565	0,5233	0,4198	0,6884	0,5086
ABCD	1	0,4706	0,5946	0,5524	0,0000	0,6429	0,4586	0,6462	0,5963	0,3810	0,6516	0,4994
Emotion Hunters	1	0,4596	0,5205	0,5842	0,2400	0,4589	0,5000	0,4319	0,5484	0,4601	0,6319	0,4835
App2Check	2	0,4048	0,3814	0,5831	0,2642	0,3614	0,5463	0,3465	0,5181	0,1250	0,2108	0,3741
App2Check	1	0,3529	0,4149	0,3855	0,4000	0,6122	0,4867	0,5340	0,4339	0,3293	0,5741	0,4523
baseline_TFIDF		0,2945	0,4444	0,4680	0,3684	0,3493	0,3314	0,5392	0,3360	0,3486	0,5944	0,4074
baseline_OHE		0,2178	0,4221	0,3526	0,2593	0,2918	0,3032	0,4564	0,3191	0,2158	0,5243	0,3362
baseline_random		0,1039	0,1541	0,2683	0,0304	0,1760	0,1529	0,2941	0,1565	0,2013	0,3426	0,1872

```
Input "NE VOGLIO ANCORA #AchilleLauroExpress
↪  #pechinoexpress https://t.co/p9O3ge86i4"
↪  Output:love,trust
Input "RT @user: #UnPassoDalCielo5 stasera x il gran
↪  finale di stagione mi sono munita di Nutella
↪  biscuits" Output:anticipation,love,trust
Input "Mi manca Nicole #IlCollegio" Output:sadness
Input "Sono stata alla Risiera di San Sabba... ho
↪  ancora i brividi al solo pensiero.. #Ulisse"
↪  Output:fear
```

3. Results

The results in the following Tables 1 and 2 suggest that our approach, even if it is not the best model of the challenge, at least generalizes quite well to a domain different from the training one. This is for us a good result since we aim at applying the model in contexts different from social media analysis. In particular, we are working on the multimodal analysis of human communication with conversational agents, in which the analysis of the textual part of verbal communication can be very important to fully understand the emotional state of the user. We are actually exploring the performances of models based on LLMs by fine-tuning the most popular one on a dataset of text denoting emotion expression not taken from tweets, which is more in line with our final goal.

References

[1] O. Araque, S. Frenda, R. Sprugnoli, D. Nozza, V. Patti, EMit at EVALITA 2023: Overview of the Categorical Emotion Detection in Italian Social Media Task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[2] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[3] R. Plutchik, 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 (2001) 344–350.

[4] M. Polignano, P. Basile, M. de Gemmis, G. Semeraro, V. Basile, Alberto: Italian BERT language understanding model for NLP challenging tasks based on tweets, in: Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy,

Table 2
F1-score evaluation for out-domain testing

team	run id	anger	anticipation	disgust	fear	joy	love	neutral	sadness	surprise	trust	macro-avg
extremITA	2	0,4051	0,4923	0,6684	0,0000	0,4416	0,7552	0,6355	0,3049	0,4138	0,8603	0,4977
extremITA	1	0,5027	0,3667	0,6219	0,0000	0,3176	0,7273	0,5634	0,2024	0,3350	0,8545	0,4491
Emotion Hunters	1	0,3671	0,6053	0,6364	0,0000	0,3768	0,5907	0,6250	0,3103	0,3692	0,8632	0,4744
App2Check	2	0,6379	0,3256	0,6790	0,1818	0,2545	0,6381	0,2564	0,3195	0,1373	0,3001	0,3730
App2Check	1	0,2710	0,4301	0,4691	0,0000	0,4167	0,6528	0,3448	0,2653	0,3662	0,8064	0,4022
baseline_TFIDF		0,3972	0,4533	0,4197	0,0000	0,1890	0,3218	0,1974	0,3129	0,2812	0,7468	0,3319
baseline_OHE		0,3342	0,4412	0,4092	0,0000	0,1786	0,3034	0,1778	0,2649	0,1913	0,6869	0,2987
baseline_random		0,1984	0,0917	0,2188	0,0081	0,1624	0,2208	0,0841	0,2097	0,1336	0,5483	0,1876

November 13-15, 2019, 2019.

[5] Openai - gpt 3.5 api large language model, 2023. URL: https://chat.openai.com/chat.

[6] R. Sprugnoli, Multiemotions-it: a new dataset for opinion polarity and emotion analysis for italian, in: Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021, 2020.

[7] L. C. Passaro, A. Lenci, Evaluating context selection strategies to build emotive vector space models, in: N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016, European Language Resources Association (ELRA), 2016.

[8] M. Pota, M. Ventura, R. Catelli, M. Esposito, An effective bert-based pipeline for twitter sentiment analysis: A case study in italian, Sensors 21 (2021) 133. doi:10.3390/s21010133.

[9] Mdz digital library team, «bert xxl italian models», hugging face, 2020. URL: https://huggingface.co/dbmdz/bert-base-italian-xxl-cased.

[10] L. Parisi, S. Francia, P. Magnani, Umberto: an italian language model trained with whole word masking, Original-date 55 (2020) 31Z.

[11] F. Bianchi, D. Nozza, D. Hovy, et al., Feel-it: Emotion and sentiment classification for the italian language, in: Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Association for Computational Linguistics, 2021.

[12] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[13] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, S. Ravi, Goemotions: A dataset of fine-grained emotions, arXiv preprint arXiv:2005.00547 (2020).

[14] M. Pota, M. Ventura, R. Catelli, M. Esposito, An effective bert-based pipeline for twitter sentiment analysis: A case study in italian, Sensors 21 (2020) 133.

[15] Emoji list, last consulted May 2023. URL: https://unicode.org/emoji/charts/full-emoji-list.html.

App2Check at EMit: Large Language Models for Multilabel Emotion Classification

Gioele Cageggi[1], Emanuele Di Rosa[2] and Asia Uboldi[3]

[1]Data Scientist at App2Check srl, Via XX Settembre, 14 - 16121, Genoa, Italy
[2]Chief Technology Officer at App2Check srl, Via XX Settembre, 14 - 16121, Genoa, Italy
[3]Data Scientist at App2Check srl, Via XX Settembre, 14 - 16121, Genoa, Italy

Abstract

In this paper we compare the performance of three state-of-the-art LLM-based approaches for multilabel emotion classification: fine-tuned multilingual T5 and two few shot prompting approaches: plain FLAN and ChatGPT. In our experimental analysis we show that FLAN T5 is the worst performer and our fine-tuned MT5 is the best performer in our dev set and, overall, is better than ChatGPT3.5 on the test set of the competition. Moreover, we show that MT5 and ChatGPT3.5 have complementary performance on different emotions and that A2C-best, our unsubmitted system that combines our best performer models for each emotion, has a macro F1 that is 0.02 greater than the winner of the competition in the out-of-domain benchmark. Finally, we suggest that a perspectivist approach is more suitable for evaluating systems on emotion detection.

Keywords

Emotion Detection, Large Language Model, ChatGPT, FLAN, mT5, Prompt Engineering

1. Introduction

Categorical Emotions Detection refers to the machine learning task of detecting the presence of specific emotions in a text. Detecting customers emotions, for example, is a useful task having many practical applications in industry, from customer experience analysis to customer churn prevention.

The categories of emotions used may vary. In this paper we consider the 8 main emotions of Plutchik's wheel [2] (anger, expectation, disgust, fear, joy, sadness, surprise, trust), plus the emotion "love," which is one of the dyads, according to the Emit 2023 competition [3], and Neutral, which is absence of emotions.

In this paper, we:

1. present three approaches for detecting emotions in a text, all based on large language models (LLM)
2. show that, on the dev set, FLAN T5 is the worst performer and our fine-tuned MT5 is the best performer
3. overall, between our models, MT5 is better than ChatGPT3.5 on the dev and test set of the competition
4. show that MT5 and ChatGPT3.5 show complementary performance on different emotions
5. present A2C-best, the unsubmitted system that combines our best performer model for each emotion. A2C-best shows a macro F1 that is 0.02 greater than the winner of the competition in the out-of-domain benchmark
6. perform error analysis on instances where all systems disagree and re-annotate them, also showing that we disagree with some labels in the golden standard
7. suggest that a "perspectivist approach" [4] is more suitable for evaluating emotions detection systems.

This paper is structured as follows: after the introduction, we describe the three approaches taken into account, then, we present and discuss the results on both dev and test set of the competition. Finally, we present our conclusions.

2. Approaches Adopted

In this paper, we study two different approaches to solve the Categorical Emotion Detection task, both Transformer-based:

- LLM Fine-tuning: Starting from a pre-trained LLM model, we use the competition dataset to fine-tune the model in order to solve the specific task
- Few-Shot Prompting: Using an Instruction Tuned LLM, prompts are designed to properly guide the model in defining its behavior for the task.

Briefly, the main differences of these two approaches are:

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT [1]
gioele.cageggi@app2check.com (G. Cageggi);
emanuele.dirosa@app2check.com (E. Di Rosa);
asia.uboldi@app2check.com (A. Uboldi)
https://www.linkedin.com/in/emanueledirosa/ (E. Di Rosa)
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper3.pdf

- While fine-tuned models require a larger labeled dataset for training, prompt-based models work even with a smaller few-shot dataset
- Fine-tuning requires high computational and resource capacity to complete the training. Few-shot prompting focuses on the refinement of prompts and instructions without changing the model parameters
- The carbon footprint of the two approaches is quite different. Fine-tuning an LLM can be computationally expensive and energy-intensive. The environmental impact generally tends to be more energy-demanding than prompt tuning, which is considered more eco-friendly because it avoids a full-scale fine-tuning process
- Fine-tuned models achieve better accuracy values when there is an abundance of labeled data, while prompt tuning can offer reasonable performance even with a limited amount of labeled data
- While fine-tuned models make LLMs specialized for a specific task, prompt tuning allows for a more flexible approach to solving different tasks with minimal changes to the prompt.

Moreover, as internal reference, we build a system called A2C-Baseline. It combines multiple ML models, such as Decision Trees [5] and KNN models [6], where we select for each emotion the best one from a pool of models. The input text is vectorized using the tf-idf methodology.

Finally, we define a voting system, A2C-Voting, that combines the prediction of each sentence from A2C-mT5-r1, A2C-GPT-r2 and A2C-Baseline. It chooses for each prediction the result with the largest agreement. The majority is always guaranteed, being based on a binary ranking of individual emotions (present/not present) and a voting system on three different predictions.

2.1. Fine-tuned LLM

Fine tuning LLMs has been proved to be an effective approach for text classification problems and in [7] we showed to be the winner approach in all tasks of the ABSITA competition. MT5 [8] is the LLM we decide to use here. It is a multilingual variant of T5 [9], a text-to-text model released by Google in 2021. T5 uses a transformer-based architecture and can be fine-tuned to return text labels for classification tasks. MT5 has been pre-trained on mC4, which is a version of Common Crawl's multilingual web crawl corpus containing 101 languages. This enables the exploitation of the potential of the T5 model on a task involving Italian text.

In this paper, in order to use this model, we use the Hugging Face API [10] wrapped by the Simple Transformers [11] library. From the available models, we choose the google/mt5-base version, which has 580 million parameters. We tried to apply google/mt5-xxl, but out of memory exception prevents us from using it in a Google Colaboratory cloud environment. More specifically, has been trained using an Nvidia A100 GPU with 40GB of memory. Training is performed for 20 epochs on 90% of the competition training dataset with a stratified split strategy. In this paper this model is referred to as A2C-mT5-r1.

2.2. Plain FLAN

FLAN-T5 [12] is one of two Few-Shot Prompting approaches that we experiment with in this paper.

It is a model based on T5[9], on which we perform instruction fine-tuning. This process entails training the model using an instruction set that describes how to perform over 1000 additional tasks. The instruction fine-tuning process involves providing the model with an instruction set and executing the tasks specified in the instructions.

In this paper, we use Hugging Face's transformers library to import the google/flan-t5-xl model and use it. Then, through prompt engineering techniques, we develop a prompt to associate an input text with one or more emotions. In the first iteration of the solution, we use a single prompt to identify and associate all possible emotions if present in the input text. However, the model is not supporting this compact approach. Thus, we modify the prompt to identify a single emotion at a time. We find better outputs with this last approach. Then we develop ten prompts, one per emotion.

The prompts start with *Detect if the text provided contains EmotionX as emotion. If the emotion is available in the input text, the value will be 1; 0 otherwise* , where *EmotionX* is the emotion to look for. Then two sentences follow, one of which contains the emotion and the other does not. In this paper this model is referred to as A2C-FlanT5.

2.3. ChatGPT

ChatGPT 3.5 is the second of the two Few-Shot Prompting approaches that we apply to experiment with in this paper. The version we use in this model is gpt-3.5-turbo-0301 [13]. The specifics of the model have not been publicly disclosed yet. It is a similar model to the previous GPT-3 model [13], trained on a set of text and code created before Q4 2021. It is then trained using a reinforcement learning method with rewards derived from human comparison.

In this paper, we use the OpenAI library [14] to process requests to the model. Unlike the approach chosen for FLAN T5, we develop a prompt to simultaneously identify all emotions for each text input. We prepared

Table 1
Example of sentences where the systems and Human classification disagree with the Golden Standard.

Text	A2C Team	Gold	A2C-mT5-r1	A2C-GPT-r2
Mi bastano 5", ma se le esibizioni di Pannofino e Facciolini totalizzano più di 20 punti possiamo annullare l'edizione... #taleequaleshow	Disgust	Anticipation	Trust	Anger
RT @user: Ultimo attacca i giornalisti in sala stampa: "Me l'avete tirata". Clima tesissimo. #Sanremo2019	Anger	Neutral	Trust	Anger
Perché lo AMano #IMedici #IMedici3	Love	Neutral	Love	Love

a prompt with six examples of text inputs, taken from the competition training dataset. All emotions have been mapped within the text examples. The output requested within the prompt is structured as a JSON with as many keys as emotions, with a value of 1 if a given emotion is present, 0 otherwise.

The prompt used is the following: *Determin the emotions in the text provided, which is delimited by <>. The available emotions are: Anger, Anticipation, Disgust, Fear, Joy, Love, Neutral, Sadness, Surprise, Trust. Provide the answer in JSON format, with the following keys: Anger, Anticipation, Disgust, Fear, Joy, Love, Neutral, Sadness, Surprise, Trust. If that emotion is present inside the input text, the value will be 1; 0 otherwise.* A series of examples then follow, in the format: *Text: <...>Answer: {"Anger":0, "Anticipation":0, "Disgust":0, "Fear":0, "Joy":0, "Love":0, "Neutral":0, "Sadness":0, "Surprise":1, "Trust":1}*

Note that this model allows to identify all emotions simultaneously, unlike FLAN T5, in which emotions have been identified one at a time. In this paper, this model is referred to as A2C-GPT-r2.

2.4. Description of our best approach: A2C-best

A2C-mT5-r1 and A2C-GPT-r2 show to be complementary in their ability to accurately detect emotions in the evaluation sets. Specifically, in the dev set, A2C-mT5-r1 outperforms A2C-GPT-r2, while the latter exhibits better performance on Anger, Disgust, Fear, and Sadness. Based on these findings, in the following, we show A2C-best, which combines the top-performing A2C models for each individual emotion.

We show in 3.1 and 3.2 the results of its application ranking on the test set of the competition as unsubmitted system, since we believe that its results are interesting for the research community.

3. Experimental Analysis

In this paper, we refer to two types of datasets: the development dataset and the competition test set. The development dataset is used to select the best A2C models to submit to the competition, while the competition test set consists of both in-domain and out-of-domain data. The dev set is split from the competition training set using the stratified technique [15], which ensures that the original proportions of labels is maintained in each subset. The training is made on the 80% of the training dataset; models are selected on the 10% of the dataset and tested on the remaining 10%. Once models to submit have been selected, we retrained them on the 100% of the training data. From here on we will refer to *Dev set* as the model selection set; *In-domain test set* and *Out-of-domain test set* refer, respectively, as the in-domain and out-of-domain competition datasets.

Tables 2 and 3 show the A2C models that participated in the competition applied on the Dev set, but also additional models developed post-deadline, highlighted in italics for a fair detection. Tables 6 and 7 include all models from both A2C and other competitors applied on the competition test set. All tables display the Macro F1 and F1 metrics for individual emotions across all models.

3.1. Results on Dev Set

In Table 2 and 3, we show the results of our model on the Dev set, where unsubmitted models are shown in italic. The worst performer is A2C-FlanT5 with an MF1 of 0.27: it shows the worst performance on the Neutral label, with an F1 score of 0. The best performer between the models we evaluated for the submission is A2C-mT5-r1, with an MF1 of 0.45, showcasing better performance on 6 out of 10 emotions when compared to the models that are not highlighted in italic. For the second run, we decide to select A2C-GPT-r2 instead of A2C-Baseline,

since it performs in a complementary way compared to A2C-mT5-r1, and to pursue a more innovative approach. More specifically, it is clear that A2C-mT5-r1 and A2C-GPT-r2 exhibit complementary performance on different emotions: A2C-GPT-r2 excels in Anger, Disgust, and Sadness, while A2C-mT5-r1 performs better in Anticipation, Joy, Neutral, Surprise, and Trust. This complementary performance is almost entirely preserved in the competition test sets as well. Based on this observation, we synthesize a post-deadline system called A2C-best which selects the model with the best performance for each emotion.

3.2. Results on Competition Test Sets

In domain test set In Tables 4 and 5, we compare both competitors systems and all our models on the in-domain Test Set of the competition. When we look at the individual emotions, ExtremITA run 2 achieves almost always the highest scores, except for Joy, where ABCD run 1 is the best one, and Love, where A2C-GPT-r2 is the best performer.

We also include in the tables the results obtained by A2C-best, which ranks second after ExtremITA's solutions, with an MF1 score at a distance of 0.005 from its first run.

The complementarity observed in the dev set between A2C-mT5-r1 and A2C-GPT-r2 also holds true within this test set, except for the emotion of Fear. We include an upper bound benchmark, Best-All, to define the potential margin of improvement by combining all the competition models.

Out-of-domain test set In Tables 6 and 7, we show the results of our systems and the other participants on the out-of-domain Test Set of the competition. Observing individual emotions, A2C-GPT-r2 shows the best score on Anger, Disgust, and Fear, while A2C-Voting on Sadness.

We obtain A2C-best by incorporating the best results of our models into one system, that selects the best of our models for each emotion. A2C-best shows to be the top performer among the submitted runs, improving the winner by 0.02 of MF1.

Once again, complementarity on emotions is clear between A2C-mT5-r1 and A2C-GPT-r2, except for Love. As an upper bound, Best-All shows that the potential margin of improvement is more significant in the out-of-domain test set.

3.3. Error analysis

In order to improve our systems performance, we randomly selected instances in which all systems disagree to analyze the most difficult cases. However, during our error analysis, we noticed that many times we did not agree on the samples annotation. In table 1, we show just 3 samples (out of many) in which we disagree with the golden standard (two different people plus a referee). The goal is to highlight whether disagreement between the systems is due to just systems that cannot correctly meet the ground truth or if such instances may be interpreted in multiple ways and thus requiring multiple, equally correct, labeling. As we can see in table 1, there are differences between the Golden Standard (Gold column) and our classification (A2C team column). The research community is working towards the direction of perspectivist approaches (see [16] and [17]) in which, well-known issues of having just one single ground truth are taken into account especially in Natural Language

Table 2
Dev set - performance of A2C models on emotions Anger, Anticipation, Disgust, Fear, Joy

MF1	Model	Ang	Ant	Dis	Fea	Joy
0.529	***A2C-best***	**0.47**	**0.60**	**0.55**	**0.40**	**0.52**
0.501	*A2C-Voting*	**0.47**	**0.60**	0.48	**0.40**	**0.52**
0.494	*A2C-r1r2*	0.41	0.56	**0.55**	0.29	0.45
0.447	A2C-mT5-r1	0.35	0.56	0.41	0.13	0.45
0.379	A2C-Baseline	0.34	0.52	0.37	0.31	0.29
0.360	A2C-GPT-r2	0.41	0.38	**0.55**	0.29	0.39
0.267	A2C-FlanT5	0.20	0.38	0.36	0.17	0.35

Table 3
Dev set - performance of A2C models on emotions Love, Neutral, Sadness, Surprise, Trust

MF1	Model	Lov	Neu	Sad	Sur	Tru
0.529	***A2C-best***	**0.55**	**0.55**	**0.54**	**0.48**	**0.62**
0.501	*A2C-Voting*	**0.55**	0.50	0.51	0.36	**0.62**
0.494	*A2C-r1r2*	0.51	0.55	**0.54**	**0.48**	0.59
0.447	A2C-mT5-r1	0.51	**0.55**	0.42	**0.48**	0.59
0.379	A2C-Baseline	0.34	0.36	0.38	0.32	0.57
0.360	A2C-GPT-r2	0.51	0.16	**0.54**	0.12	0.25
0.267	A2C-FlanT5	0.42	0.00	0.36	0.28	0.15

Table 4
In-domain Test set - performance of all systems on emotions Anger, Anticipation, Disgust, Fear, Joy

MF1	Model	Ang	Ant	Dis	Fea	Joy
0.608	*Best-All*	**0.52**	**0.64**	**0.63**	**0.58**	**0.64**
0.603	**extremITA2**	**0.52**	**0.64**	**0.63**	**0.58**	0.62
0.509	extremITA1	0.48	0.56	0.57	0.14	0.59
0.504	*A2C-best*	0.44	0.49	0.58	0.40	0.61
0.499	ABCD1	0.47	0.59	0.55	0.00	**0.64**
0.492	*A2C-r1r2*	0.40	0.41	0.58	0.40	0.61
0.484	E.Hunters1	0.46	0.52	0.58	0.24	0.46
0.452	A2C-mT5-r1	0.35	0.41	0.39	0.40	0.61
0.445	*A2C-Voting*	0.44	0.49	0.44	0.24	0.55
0.374	A2C-GPT-r2	0.40	0.38	0.58	0.26	0.36
0.329	A2C-FlanT5	0.25	0.34	0.38	0.33	0.34
0.299	A2C-Baseline	0.28	0.40	0.30	0.10	0.12

Table 5
In-domain Test set - performance of all systems on emotions Love, Neutral, Sadness, Surprise, Trust

MF1	Model	Lov	Neu	Sad	Sur	Tru
0.608	*Best-All*	**0.55**	**0.70**	**0.63**	**0.51**	**0.69**
0.603	**extremITA2**	0.52	**0.70**	**0.63**	**0.51**	**0.69**
0.509	extremITA1	0.45	0.66	0.52	0.42	**0.69**
0.504	*A2C-best*	**0.55**	0.53	0.52	0.33	0.58
0.499	ABCD1	0.46	0.65	0.60	0.38	0.65
0.492	*A2C-r1r2*	**0.55**	0.53	0.52	0.33	0.57
0.484	E.Hunters1	0.50	0.43	0.55	0.46	0.63
0.452	A2C-mT5-r1	0.49	0.53	0.43	0.33	0.57
0.445	*A2C-Voting*	0.54	0.51	0.42	0.24	0.58
0.374	A2C-GPT-r2	**0.55**	0.35	0.52	0.13	0.21
0.329	A2C-FlanT5	0.38	0.48	0.34	0.31	0.15
0.299	A2C-Baseline	0.34	0.40	0.20	0.30	0.55

Table 6
Out-of-domain Test set - performance of all systems on emotions Anger, Anticipation, Disgust, Fear, Joy

MF1	Model	Ang	Ant	Dis	Fea	Joy
0.564	*Best-All*	**0.64**	**0.60**	**0.68**	**0.18**	**0.44**
0.518	***A2C-best***	**0.64**	**0.60**	**0.68**	**0.18**	0.42
0.498	extremITA2	0.41	0.49	0.67	0.00	**0.44**
0.484	*A2C-r1r2*	**0.64**	0.43	**0.68**	**0.18**	0.42
0.449	extremITA1	0.50	0.37	0.62	0.00	0.32
0.438	*A2C-Voting*	0.39	**0.60**	0.65	0.00	0.25
0.402	A2C-mT5-r1	0.27	0.43	0.47	0.00	0.42
0.373	A2C-GPT-r2	**0.64**	0.33	**0.68**	**0.18**	0.25
0.303	A2C-Baseline	0.23	0.45	0.46	0.00	0.14
0.295	A2C-FlanT5	0.51	0.22	0.59	0.00	0.26

Table 7
Out-of-domain Test set - performance of all systems on emotions Love, Neutral, Sadness, Surprise, Trust

MF1	Model	Lov	Neu	Sad	Sur	Tru
0.564	*Best-All*	**0.76**	**0.64**	**0.44**	**0.41**	**0.86**
0.518	***A2C-best***	0.71	0.34	**0.44**	0.37	0.81
0.498	extremITA2	**0.76**	**0.64**	0.30	**0.41**	**0.86**
0.484	*A2C-r1r2*	0.65	0.34	0.32	0.37	0.81
0.449	extremITA1	0.73	0.56	0.20	0.34	0.85
0.438	*A2C-Voting*	0.71	0.31	**0.44**	0.24	0.79
0.402	A2C-mT5-r1	0.65	0.34	0.27	0.37	0.81
0.373	A2C-GPT-r2	0.64	0.26	0.32	0.14	0.30
0.303	A2C-Baseline	0.24	0.14	0.40	0.22	0.77
0.295	A2C-FlanT5	0.39	0.27	0.27	0.19	0.25

Processing (NLP), and propose multiple equally correct labeling samples. In our opinion, categorical emotions detection is one relevant example of NLP in which is very difficult to agree on just one golden standard.

4. Conclusion

In this paper we presented the systems runs we submitted at Emit 2023 competition for emotion detection in text, and also our post deadline system called A2C-best. In particular, we presented the performance of three different LLM-based approaches, such as fine-tuned multilingual T5, and two few shot prompting techniques, A2C-GPT-r2 and FLAN T5. Our A2C-best model shows significant improvement to our official run and comparable performance with the first ranker of the competition in the out-of domain run. A2C-best scores 0.099 below the winner in the in-domain run. Finally, after relabeling difficult instances where all systems and humans disagree, we suggested that a perspectivist approach is more suitable for evaluating systems on emotion detection.

References

[1] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[2] K. K. Imbir, Psychoevolutionary Theory of Emotion (Plutchik), Springer International Publishing, Cham, 2017, pp. 1–9. URL: https://doi.org/10.1007/978-3-319-28099-8_547-1. doi:10.1007/978-3-319-28099-8_547-1.

[3] O. Araque, S. Frenda, R. Sprugnoli, D. Nozza, V. Patti, EMit at EVALITA 2023: Overview of the Categorical Emotion Detection in Italian Social Media Task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[4] V. Basile, F. Cabitza, A. Campagner, M. Fell, Toward a perspectivist turn in ground truthing for predictive computing, CoRR abs/2109.04270 (2021). URL: https://arxiv.org/abs/2109.04270. arXiv:2109.04270.

[5] L. Rokach, O. Maimon, Top-down induction of decision trees classifiers - a survey, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 35 (2005) 476–487. doi:10.1109/TSMCC.2004.843247.

[6] C. D. M. et al., Introduction to information retrieval, Cambridge University Press, 2008. URL: https://nlp.stanford.edu/IR-book/pdf/irbookprint.pdf. doi:10.1017/CBO9780511809071.

[7] E. D. Rosa, A. Durante, App2check @ ate_absita

2020: Aspect term extraction and aspect-based sentiment analysis, in: V. Basile, D. Croce, M. D. Maro, L. C. Passaro (Eds.), Proceedings of (EVALITA 2020), volume 2765 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: https://ceur-ws.org/Vol-2765/paper122.pdf.

[8] L. X. et al., mt5: A massively multilingual pre-trained text-to-text transformer, in: K. T. et al. (Ed.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, Associa tion for Computational Linguistics, 2021, pp. 483–498. URL: https://doi.org/10.18653/v1/2021.naacl-main.41. doi:10.18653/v1/2021.naacl-main.41.

[9] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, 2020. arXiv:1910.10683.

[10] Hugging Face website, 2023. URL: https://huggingface.co/.

[11] T. C. Rajapakse, Simple transformers, https://github.com/ThilinaRajapakse/simpletransformers, 2019.

[12] H. W. C. et al., Scaling instruction-finetuned language models, 2022. arXiv:2210.11416.

[13] L. O. et al., Training language models to follow instructions with human feedback, 2022. arXiv:2203.02155.

[14] Openai website, 2023. URL: https://openai.com/.

[15] K. Sechidis, G. Tsoumakas, I. P. Vlahavas, On the stratification of multi-label data, in: ECML/PKDD, 2011.

[16] F. Cabitza, , A. Campagner, V. Basile, Toward a perspectivist turn in ground truthing for predictive computing, Washington DC, USA, 2023.

[17] The perspectivist data manifesto, 2023. URL: https://pdai.info/.

ABCD team at EMit: Ensemble Approach for Categorial Emotion Detection in Social Media Messages

Nguyen Ba Dai[1,3,4], Nguyen Ngoc Phuong Uyen[2,3,5] and Dang Van Thin[1,3]

[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]University of Economics and Law, Ho Chi Minh City, Vietnam
[3]Vietnam National University, Ho Chi Minh City, Vietnam
[4]Faculty of Information Science and Engineering, University of Information Technology, Ho Chi Minh City, Vietnam
[5]Faculty of Information Systems, University of Economics and Law, Ho Chi Minh City, Vietnam

Abstract

This paper presents a submission system for the EMit (Emotions in Italian) shared task at EVALITA 2023, which focuses on categorial emotion detection in social media messages related to TV shows, TV series, music videos, and advertisements domain. We employ an ensemble approach, leveraging the BERT (Bidirectional Encoder Representations from Transformers) models known for its advanced language understanding capabilities. The BERT model is fine-tuned using domain-specific data to enhance its performance in emotion detection. The ensemble architecture combines multiple pre-trained models and utilizes a soft voting technique for robust decision-making. The results demonstrate the effectiveness of the team's ensemble model, achieving a Top 3 ranking in task 1 with 49.94% of the F1-score.

Keywords

Ensemble Approach, Categorical Emotion Detection, BERT, Sentiment Analysis, NLP, Ensemble Model

1. Introduction

Within several fields, including Natural Language Processing (NLP), Artificial Intelligence (AI) has made significant contributions in offering efficient answers for crucial societal and human issues. Two critical areas of NLP are sentiment analysis and emotional recognition [1]. While sentiment analysis usually classifies data into three main categories (positive, negative, neutral), emotional recognition extracts distinct human emotions such as disgust, fear, joy, and more. Emotion detection has been studied and applied extensively in computational and linguistic techniques to help computers understand and, at times, generate human languages. We can understand its significance because emotions play vital roles in the existence or the complete make-up of individuals. EVALITA 2023 [2] provides a shared framework for the evaluation of different systems and approaches on the EMit (Emotions in Italian) shared task [3]. In this shared task, two sub-tasks, both designed as multilabel classification problems, were proposed for participants. The first challenge called Categorial Emotion Detection, aims to identify emotions in social media messages or the absence of emotions. The EMit of EVALITA offered 8 emotional labels defined by Plutchik (anger, anticipation, disgust, fear, joy, sadness, surprise, trust, and love) [4]. In this sub-task, we detect emotions from data provided by task organizers in the social media domain related to TV shows, TV series, music videos, and advertisements. The study of emotions in social media, employed as a social signal for capturing the emotional reactions of the Italian audience, can offer artists or broadcasters highly valuable finer-grained information in assessing the delivered content.

2. Related Work

Detection of emotions is not entirely novel, for more than a decade, there have been several international evaluation campaigns launched for example the Emotion Classification shared task at WASSA 2022 [5], EmoContext at SemEval 2019 [6], The Affective Text shared task at SemEval 2007 [7],... However, further research is still needed, especially on identifying emotions through messages related to TV shows, TV series, music videos, and advertisements.

In EVALITA 2018, the Hate Speech Detection (HaSpeeDe) task was conducted specifically for Italian, with the objective of automatically labeling messages from Twitter and Facebook as either containing or not containing hate speech. To develop robust hate speech detection systems, [8] utilized the dataset released for the HaSpeeDe shared task, which combined an English dataset and a German dataset distributed for the Identification of Offen-

EVALITA 2023, September 7–8, 2023, Parma, Italy
21521914@gm.uit.edu.vn (N. B. Dai); uyennnp21411@st.uel.edu.vn (N. N. P. Uyen); thindv@uit.edu.vn (D. V. Thin)
https://nlp.uit.edu.vn/ (D. V. Thin)
0009-0008-8559-3154 (N. B. Dai); 0009-0001-8604-1991 (N. N. P. Uyen); 0000-0001-8340-1405 (D. V. Thin)
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper4.pdf

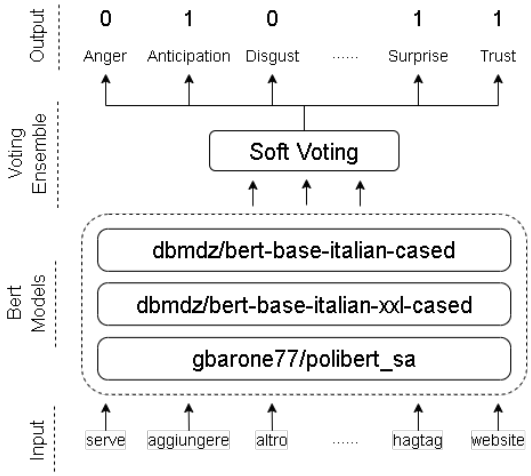


Figure 1: Overview of our ensemble model for the Task A.

sive Language shared task organized at Germeval 2018. Their findings identified a recurrent neural architecture that demonstrates stability and high performance across various languages. They also evaluated the impact of several components commonly used in the task, including the type of embeddings, the incorporation of additional features (text-based or emotion-based), the role of hashtag normalization, and the influence of emojis. [9] introduced FEEL-IT, a novel benchmark corpus of Italian Twitter posts annotated with four basic emotions: anger, fear, joy, and sadness. Another relevant research by [10], paper evaluated BERT's performance in emotion recognition using the EmotionLines dataset from the Friends Television Sitcom series and the EmotionPush dataset from Facebook Messenger chat.

Research on emotion detection has been conducted across various languages and tested on multiple technological models: an attention-based methodology for identifying and categorizing emotions in textual interactions [11], using the Bi-LSTM to classify emotions in textual and emoji utterances [12], detecting text emotions in social networks with a novel ensemble classifier based on the Parzen Tree Estimator (TPE) [13], applying K-NN and NB ML techniques in the detection of emotions [14]. To further contribute to emotion detection experiments and conduct research on identifying emotions in Italian social media audiences, we employ an ensemble approach for Categorical Emotion Detection in Social Media Messages.

3. Approach

Figure 1 shows the overview of our ensemble model for task A. The combined model integrates multiple pre-trained models to leverage their domain-specific expertise and enhance performance. By using the soft voting method, the model aggregates the predictions from each model to generate accurate results.

The integration process involves obtaining independent outputs from each pre-trained model and combining them using the soft voting method. This method considers the confidence scores associated with each prediction, assigning higher weights to more reliable predictions for a robust final decision. The model architecture facilitates seamless communication and coordination between the pre-trained models, enabling efficient information exchange and fusion of predictions. It can be easily adapted and extended to incorporate new models and emerging techniques for continuous improvement.

In this paper, we utilize the pre-trained BERT model [15] for the following reasons: (1) First, BERT is a state-of-the-art model renowned for its exceptional language understanding capabilities. It effectively captures the semantics and nuances of text due to its deep contextualized representations. (2) Second, by leveraging transfer learning, BERT provides a significant advantage. It has been pre-trained on extensive and diverse datasets, allowing us to benefit from the knowledge it has acquired. This saves computational resources and time compared to training a model from scratch. (3) Finally, BERT's adaptability to different domains is a key factor. Its pre-training covers a wide range of domains, making it flexible for various tasks. By fine-tuning BERT with domain-specific data, we can enhance its performance for our specific task.

Table 1
The general information of official datasets in our experiments.

Information	Training set	Test set
Number of samples	6 359	1 000
Number of tokens	93 621	15 713
Number of unique targets	23 028	6 215
The average length	102	102
The maximum length	304	302

Table 2
Performance Rankings of Teams in F1-Score.

Ranking	Team	Run ID	F1-Score
Top 1	extremITA	2	60.28
Top 2	extremITA	1	50.86
Top 4	EmotionHunters	1	48.35
Top 5	App2Check	2	37.41
Ours (Top 3)	**ABCD Team**	**1**	**49.94**

4. Experimental Setup

4.1. Dataset

We only use the official training set [3], which is provided by the organizers, to train our models for task A in the shared task.

Table 1 presents the general statistics of the training and testing set. As depicted in Table 1, it is evident that the training and test sets in this shared-task experiment are balanced in terms of average length. These statistics provide a foundation for understanding the data and designing appropriate models for the given task.

4.2. Pre-processing

Pre-processing steps are essential in classification-type tasks to improve the quality of the data and facilitate effective analysis. Additionally, the provided dataset is collected from social media messages, therefore, it is necessary to design the list of pre-processing steps. Based on the analysis of the training dataset, we design the list of pre-processing steps in our system as follows:

- **Step 1**: We removed the "@USER" and the placeholder "_" carried after it
- **Step 2**: We removed the redundant symbolic expressions and keep only one symbol, for example, "!!!", "???" transformed to "!", "?".
- **Step 3**: We transformed the emoji from the dataset to text because of the large amount of emoji in sentences.
- **Step 4**: We tagged the link, phone number, and hashtag to its related token, for example, "0123456789" transformed to "<phone>".
- **Step 5**: Finally, we removed the extra space symbols from the text.

4.3. Models

After investigating various models, we have concluded that the dbmdz/bert-base-italian-cased, gbarone77/polibert_sa, and dbmdz/bert-base-italian-xxl-cased BERT models are the most suitable choices for this shared task. We have identified several reasons to support our decision: (1) These models were trained on the Italia corpus, which aligns well with the requirements of this shared task. (2) All three models have demonstrated outstanding performance for this task, as indicated in Table 3.

4.4. System setting

We evaluated various models using the HuggingFace Transformer library [16]. Each model was trained with a fixed number of epochs, specifically 5 epochs. The learning rate used for dbmdz/bert-base-italian-cased and gbarone77/polibert_sa was set to 3e-5, while for dbmdz/bert-base-italian-xxl-cased, it was set to 2e-5. The batch size for all pre-trained language models was set to 16. No development data was utilized for model tuning. We employed the AdamW optimizer to optimize our models. Additionally, we added an extra sigmoid layer at the output for each model. To ensure reproducibility, we set a fixed random seed of 42 for training the models.

5. Main results

The official results and the results of the top systems are shown in Table 2. Our best model achieves the Top 3 ranking in the challenge during the final round. Our model achieved a result of 49.94% in terms of F1-score,

Table 3
Comparison of Models and F1-Score Scores.

Models	F1-Score
dbmdz/bert-base-italian-cased	**43.22**
dbmdz/bert-base-italian-xxl-cased	**47.07**
gbarone77/polibert_sa	**47.24**
bert-base-multilingual-cased	39.03
Geotrend/bert-base-it-cased	40.37
mgrella/autonlp-bank-transaction-classification-5521155	38.02
nlptown/bert-base-multilingual-uncased-sentiment	40.20
Babelscape/wikineural-multilingual-ner	38.02

which is lower than the F1 scores of the Top 1 and Top 2 teams, which are +10.34% and +0.92%, respectively.

Table 3 shows the overall results of our submission model and other variants on the test set on the challenge. Overall, it can be seen that the performance of the model is improved when we preprocessed data input, fine-tuned, and ensemble those models. These techniques, mentioned in previous work [17], have consistently shown their effectiveness in enhancing model performance. Preprocessing improves data quality, fine-tuning tailors models to the task, and ensembling combines strengths for better results. These findings reinforce the value of these techniques for improving model performance in similar domains or tasks. Our ensemble methods have significantly improved the F1-Score, increasing it from 2.7% to 6.42%. This demonstrates the effectiveness of ensembling in enhancing model performance on the task at hand.

On the other hand, Table 3 shows that the models dbmdz/bert-base-italian-cased, dbmdz/bert-base-italian-xxl-cased, and gbarone77/polibert_sa outperformed the other models. Among them, the gbarone77/polibert_sa model achieved the best performance due to its utilization of Italian language data and its specific training for sentiment analysis, which aligns well with the task requirements. The superior performance of these three models can be attributed to their training on a large amount of data specific to the Italian language.

In contrast, the models bert-base-multilingual-cased, Babelscape/wikineural-multilingual-ner, and nlptown/bert-base-multilingual-uncased-sentiment were trained on multilingual language data, which resulted in comparatively lower performance when applied to Italian-specific tasks. To address this issue, the Geotrend/bert-base-it-cased model was introduced as an enhancement to the bert-base-multilingual-cased model by training it on Italian language data. This targeted training on Italian data resulted in a performance improvement of 1.34% compared to the original multilingual model. However, despite this improvement, the Geotrend/bert-base-it-cased model still exhibits lower performance than the top three models mentioned earlier. One possible reason for this is that the amount of training data used for the Geotrend/bert-base-it-cased model was relatively smaller compared to the extensive data used for the top-performing models. The mgrella/autonlp-bank-transaction-classification-5521155 models, while trained on Italian language data, were specifically tailored for the ”bank transaction” field. As a result, their performance is lower for sentiment analysis tasks, which is the focus of our study.

6. Conclusion

In this paper, we present our submission system for the EMit at EVALITA 2023: The Categorial Emotion Detection Task, where our approach achieved a Top 3 ranking. Instead of relying on a single model, we adopt an ensemble method approach using the pre-trained BERT model [15] to tackle the task. Through extensive experimentation and analysis, we have found this approach to be highly effective and believe it can be applied to other domains for sentiment analysis as well. While we acknowledge that our familiarity with the Italian language was limited, we recognize the importance of effective text pre-processing techniques in enhancing performance on the test set. Although we were unable to leverage these techniques to their full potential, we firmly believe that incorporating appropriate text pre-processing techniques can further improve the performance of our model.

References

[1] F. A. Acheampong, C. Wenyu, H. Nunoo-Mensah, Text-based emotion detection: Advances, challenges, and opportunities, Engineering Reports 2 (2020) e12189.

[2] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th

evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[3] O. Araque, S. Frenda, R. Sprugnoli, D. Nozza, V. Patti, EMit at EVALITA 2023: Overview of the Categorical Emotion Detection in Italian Social Media Task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[4] L. A. Camras, R. Plutchik, H. Kellerman, Emotion: Theory, research, and experience. vol. 1. theories of emotion, American Journal of Psychology 94 (1981) 370.

[5] V. Barriere, S. Tafreshi, J. Sedoc, S. Alqahtani, Wassa 2022 shared task: Predicting empathy, emotion and personality in reaction to news stories, in: Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, 2022, pp. 214–227.

[6] A. Chatterjee, K. N. Narahari, M. Joshi, P. Agrawal, Semeval-2019 task 3: Emocontext contextual emotion detection in text, in: Proceedings of the 13th international workshop on semantic evaluation, 2019, pp. 39–48.

[7] E. Agirre, L. Márquez, R. Wicentowski, Proceedings of the fourth international workshop on semantic evaluations (semeval-2007), in: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), 2007.

[8] M. Corazza, S. Menini, E. Cabrio, S. Tonelli, S. Villata, A multilingual evaluation for online hate speech detection, ACM Transactions on Internet Technology (TOIT) 20 (2020) 1–22.

[9] F. Bianchi, D. Nozza, D. Hovy, et al., Feel-it: Emotion and sentiment classification for the italian language, in: Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Association for Computational Linguistics, 2021.

[10] Y.-H. Huang, S.-R. Lee, M.-Y. Ma, Y.-H. Chen, Y.-W. Yu, Y.-S. Chen, Emotionx-idea: Emotion bert–an affectional model for conversation, arXiv preprint arXiv:1908.06264 (2019).

[11] W. Ragheb, J. Azé, S. Bringay, M. Servajean, Attention-based modeling for emotion detection and classification in textual conversations, arXiv preprint arXiv:1906.07020 (2019).

[12] L. Ma, L. Zhang, W. Ye, W. Hu, Pkuse at semeval-2019 task 3: emotion detection with emotion-oriented neural attention network, in: Proceedings of the 13th international workshop on semantic evaluation, 2019, pp. 287–291.

[13] F. Ghanbari-Adivi, M. Mosleh, Text emotion detection in social networks using a novel ensemble classifier based on parzen tree estimator (tpe), Neural Computing and Applications 31 (2019) 8971–8983.

[14] M. Suhasini, B. Srinivasu, Emotion detection framework for twitter data using supervised classifiers, in: Data Engineering and Communication Technology: Proceedings of 3rd ICDECT-2K19, Springer, 2020, pp. 565–576.

[15] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[16] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6. doi:10.18653/v1/2020.emnlp-demos.6.

[17] Y. Xu, X. Qiu, L. Zhou, X. Huang, Improving bert fine-tuning via self-ensemble and self-distillation, arXiv preprint arXiv:2002.10345 (2020).

EmotivITA at EVALITA2023: Overview of the Dimensional and Multidimensional Emotion Analysis Task

Giovanni Gafà[1,*], Francesco Cutugno[2] and Marco Venuti[1]

[1]University of Catania, Italy
[2]University of Naples Federico II, Italy

Abstract

EmotivITA is the first shared task for Italian Dimensional and Multidimensional Emotion Analysis, aiming to promote research in the field of emotion detection within the Italian language. We developed an Italian dataset annotated following the dimensional model of emotions and invited participants to submit systems to predict Valence, Arousal, and Dominance associated to sentences in the *corpus*. Five runs were submitted by two teams. We present the dataset, the evaluation methodology, and the approaches of the participating systems.

Keywords

emotion analysis, emotion detection, VAD model, dataset, EmoITA, EmotivITA, Evalita 2023,

1. Introduction and Motivation

In the last two decades, the analysis of emotions that people express in texts has become an essential area in Natural Language Processing (NLP). Such an interest springs from the awareness of the crucial role feelings have in our cognition: being able to detect and eventually simulate them could be a fundamental step to produce human-like forms of artificial intelligence [1]. For a review on possible applications of Emotion Analysis (EA), ranging from stock market predictions to the management of catastrophic events, see for example [2].

Taking into account the somewhat uncertain terminology about human feelings occasionally found in the literature (see below), we start by defining some terms. Adopting a well known typology of affective states by Scherer [3, pp. 140–141], we use the word 'emotion' to refer to a "relatively brief episode of synchronized responses by all or most organismic subsystems to the evaluation of an external or internal event as being of major significance", whereas 'sentiments', like Scherer's 'attitudes', are "relatively enduring, affectively colored beliefs, preferences, and predispositions toward objects or persons".

Sentiment analysis has been a major interest for computational linguistics for a long time, and, over the years, it moved from the prediction of the semantic polarity towards more fine-grained modeling, as is the case in *Aspect-based Sentiment Analysis* [4] and *Stance Detection* [5]; similar studies have been conducted on Italian texts as well [6, 7].

Recently, EA started receiving more and more attention as well. Several models of emotion proposed in psychology have been used in NLP, either categorical or dimensional. The former consider feelings as discrete, and usually identify a small set of basic emotions upon which other, more subtle and complex affective states are built; the widely adopted model conceived by Ekman [8], for instance, proposes six fundamental emotions. The latter, on the contrary, describes emotions by combining a limited number of independent dimensions in a real-valued vector space. The model proposed by Russel and Mehrabian [9], probably the best-known, recognizes three dimensions: *Valence* (measuring pleasure or displeasure), *Arousal* (degree of excitement or calm), and *Dominance* (level of control over the situation) – the VAD model.

Categorical models have some advantages over dimensional ones, as they allow the identification of several emotions in the same input and usually have simpler interpretations. Nevertheless, they have been criticized for their use of culture and language specific labels [10]; besides, different categorical models adopt different sets of emotions, making it difficult to compare studies. Concerning dimensional models, the independence of the three dimensions is yet to be ascertained [11, 12]; however, dimensional models allow easier comparisons between emotions and can describe feelings that are difficult to label.

At SemEval, the most renowned evaluation campaign of NLP, the first shared task concerning emotion detection (for three languages: English, Arabic and Spanish) was proposed in 2018 [13]. Building on earlier works, a 22,000 tweet dataset was annotated for many different affect states, following both the categorical and dimensional models of emotions (limited to the Valence dimen-

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
*Corresponding author.
giovanni.gafa@phd.unict.it (G. Gafà); cutugno@unina.it (F. Cutugno); marco.venuti@unict.it (M. Venuti)
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper5.pdf

sion[1]); the sub-tasks involved emotion classification and emotion regression.

Another task of emotion classification was proposed at SemEval 2019 [14], this time leveraging a dataset containing roughly 3,000 short conversations annotated for the presence of four emotions; the purpose was to study and exploit the role of context in facilitating emotion detection.

Anyway, EA has not yet received in Italy the same amount of interest it gained at the international level. This is probably due to the lack of resources annotated for emotions. After some investigations, we could find just a few *lexica* [15, 16, 17]; some are not open to the public [18] or are quite specific in scope [19]; others are the result of automatic translations from English of existing vocabularies, and have not been re-annotated by Italian speakers [20, 21]. This situation worsens when it comes to datasets, where to the best of our knowledge only domain-specific resources are available [22, 23, 24]. Another dataset [25] has been proposed at Evalita 2023 [26], containing social media messages about TV shows, TV series, music videos, and advertisements, which had been labeled following the Plutchik model of emotions [27].

As we tried to outline, existing datasets for EA in Italian are scarce and quite specialized. Moreover, the emotion formats used for annotating the *corpora* are uniquely categorical. Nevertheless, dimensional models are receiving increasing attention in tasks of emotion detection [28, 29]. By proposing the EmotivITA shared task at the Evalita 2023 evaluation campaign, we aim at providing a new, general-purpose resource for EA in Italian, with labeling provided by Italian speakers, EmoITA: a dataset composed with a genre and domain-balanced selection of more than 10,000 written sentences, annotated following the dimensional model of emotions; on the other hand, we intend to promote dimensional and multidimensional EA in Italian.

The rest of the paper is organized as follows: Section 2 provides a definition of the task; Section 3 describes the dataset made available to participants, and the process of its creation; Section 4 details the official evaluation measures; Section 5 reports the results obtained by participating teams; Section 6 discusses the results; in Section 7 we draw some conclusions on the outcomes of the task.

2. Definition of the task

The EmotivITA shared task consists of automatically annotating for emotions in the VAD model a collection of written sentences from a genre-balanced dataset translated into Italian. More specifically, the task has been organized into two sub-tasks whose results will be evaluated separately:

- **Sub-task A: Dimensional emotion regression**. Prediction of Valence, Arousal, and Dominance values based on a set of Italian sentences and annotations, using only the target annotated dimension for training – so, for instance, when predicting Valence participant systems may only use Valence values annotated in the dataset for training; the same holds for Arousal and Dominance.
- **Sub-task B: Multidimensional emotion regression**. Prediction of Valence, Arousal, and Dominance values based on a set of Italian sentences and annotations, using all mentioned dimensions for training – so participant systems should determine Valence, Arousal, and Dominance simultaneously, using values from the three dimensions for training.

Both sub-tasks are regression problems, so participating teams were asked to provide in the output the sentence id and three real numbers between 1 and 5, relative to the three predicted dimensions. Sub-task B intends to study and exploit potential correlations between Valence, Arousal, and Dominance, which have been discussed in the literature (see § 1).

Participants could carry out either both sub-tasks or only one of them, even if participation in sub-task A was strongly recommended, in order to have a common basis for comparison. Each participating team was allowed to submit a maximum of 2 runs for each sub-task. All runs could be produced according to the 'constrained' or 'unconstrained' modality (or both); however, we asked to specify the type of run. In constrained modality, only annotated data distributed by the organizers could be used for training and tuning the systems. Other linguistic resources (e.g., word embeddings and lexicons) were instead allowed. In unconstrained modality, annotated external data could also be employed and had to be described in the system reports.

3. Dataset

As mentioned above, the data released for the shared task derive from the Italian translation of an existing dataset, EmoBank [30]. EmoBank is the largest genre-balanced English dataset annotated employing the VAD model of emotions. As shown in Table 1, it mainly consists of the *MASC: Manually Annotated Sub-Corpus of the American National Corpus* [31], with roughly 10% of the sentences coming from the dataset of SemEval-2007 Task 14 [32].

The 10,062 sentences were originally annotated by English native speakers according to two different perspec-

[1]As a case in point of inaccuracy when dealing with emotion-related terms, Valence was regarded as an equivalent of 'sentiment' throughout the study.

Table 1
Genre distribution of the EmoBank *corpus.*

Corpus	Domain	# Sentences
SemEval-2007	news headlines	1,192
MASC	blogs	1,336
	essays	1,135
	fiction	2,753
	letters	1,413
	newspapers	1,314
	travel guides	919
Total		10,062

tives: the emotion they felt the writer meant to express, and the emotion evoked in an average reader.

At first, the Italian version of the dataset was studied as part of a Master's degree thesis discussed in 2022 at the Department of Humanities of the University of Catania. In this context, the sentences were initially translated automatically to Italian using the neural machine translation service offered by Microsoft Azure. As we were not satisfied with the results, a manual revision was performed splitting the *corpus* evenly between nine Italian native speakers, researchers in linguistics affiliated with Interdepartmental Research Center Urban/Eco at the University of Naples Federico II.

We also conducted a pilot study asking two of the participants to independently annotate VAD values from the reader's perspective for a small sample of sentences (150). We chose the reader's perspective because, according to Buechel and Hahn, it yields better inter-annotator agreement (IAA). For annotation, we used the Self-Assessment Manikin (SAM), a pictographic scale to assess emotional response [33, 34] already adopted for EmoBank. SAM consists of three sets of anthropomorphic cartoons displaying differences in Valence, Arousal, and Dominance values, respectively as shown in Figure 1.

We asked participants to attribute a value between 1 (minimum Valence, Arousal, and Dominance) and 5 (maximum Valence, Arousal, and Dominance), with 4 intermediate steps of 0.5. This results in a 9-point scale like the one originally proposed by Bradley and Lang (Buechel and Hahn preferred a 5-point scale). Instructions were adapted from those used for EmoBank and are available for further analysis upon request.

To measure IAA we used Pearson's correlation coefficient (r) and Mean Absolute Error (*MAE*), as other metrics like Cohen's k are not designed for scale variables (see § 4). We obtained encouraging scores in both measures for all three dimensions, with an average of 0.593 for r, indicating a large effect (see Table 2). Therefore, we decided to ask all participants to annotate the remaining sentences individually (one annotator per sentence). We then used the new labeling to fine-tune several models of transformers for dimensional EA, but the scores were

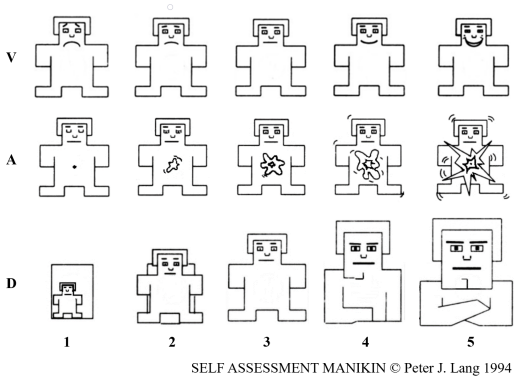


Figure 1: The SAM scales for VAD values. Dimensions (Valence, Arousal and Dominance) are reported in rows, values (from 1 to 5) in columns. Copyright of SAM by Peter J. Lang 1994.

Table 2
IAA for the three dimensions in the pilot study.

	V	A	D	Average
r	0.794	0.552	0.676	0.593
MAE	0.357	0.900	0.583	0.613

significantly worse than those obtained with the original values from the EmoBank dataset (*MAE* was between 2 and 3 times higher, r for Valence and Arousal was respectively 33% and 13% lower). This was probably due to the lack of consistency from having a single annotation for a sentence.

Moreover, we reviewed the manual revisions of the translations and found that, in at least half of the cases, the quality was still poor, either because the translated sentence did not feel natural in Italian or because it contained some kind of error.

To produce EmoITA, we resolved to start over the entire process, only keeping the approximately 5,000 translations we considered good enough. This time, we chose 16 students from the Master's Degree in Foreign Languages at the University of Catania. All of them are Italian native speakers and are specializing in English. The sentences were split among the participants: we asked to revise the 5,000 translations we kept from previous work and to propose new translations for the rest of the *corpus.* The same group of subjects also labeled each Italian sentence, and we took care never to ask a participant to annotate a sentence he had translated. Overall, we obtained 7 different annotations for each sentence, and we judge the quality of translation is now satisfactory if not perfect.

To evaluate the annotations, we proceeded similarly to the original EmoBank study: we calculated r and *MAE* between each individual series of annotations and the

Table 3
IAA for the three dimensions in EmoITA.

	V	A	D	Average
r	0.702	0.507	0.535	0.581
MAE	0.496	0.536	0.489	0.507

Table 4
Annotation examples from the development dataset.

id	text	V	A	D
260	Jet si capovolge durante una tormenta, nessun morto	3.33	4.17	3
261	Certo, risposi.	3.83	3.17	3.83

aggregated values in EmoITA, and then averaged those values for each dimension (see Table 3).

The values of *r* indicate a large effect in every dimension, particularly for Valence. Correlation is a little higher in Dominance than in Arousal, as per our pilot study: this is somewhat unusual, as in most research we analyzed regarding the English language the opposite is true. *MAE* is not as good, but still acceptable (10% of the 1-5 scale). Overall, scores are in line with those of EmoBank (*r*= 0.634 and *MAE*= 0.386, on average). They could probably get better analyzing outliers and excluding some of the annotations whose disagreement is particularly strong, a process we have not yet started at this time.

For the shared task the dataset was randomly split into a development and a test set of 8,000 and 2,062 sentences respectively (79.5% and 20.5%), taking care to preserve the genre distribution in the *corpus* (with a 1% tolerance). The development set was provided as a UTF-8, CSV comma-separated file, reporting the following fields:

id, text, V, A, D

where:

1. 'id' denotes the unique identifier of the sentence
2. 'text' denotes the text of the sentence
3. 'V' denotes the average Valence value annotated for the sentence.
4. 'A' denotes the average Arousal value annotated for the sentence.
5. 'D' denotes the average Dominance value annotated for the sentence.

See Table 4 for a couple of examples.

The test set followed the same format, but labels for Valence, Arousal and Domination were not provided.

4. Evaluation Measures

The two sub-tasks are evaluated separately comparing results obtained by participant systems with the gold standard annotations of the test set. Both constrained and unconstrained runs for a sub-task are reported in the same ranking, but we specify the type of the run.

Evaluation metrics for both sub-tasks are the standard metrics known in the literature for emotion regression that we already mentioned throughout this paper: we measure IAA based on *r* and *MAE*. The first metric estimates linear dependence between two series of data points: $\mathbf{x} = x_1, ..., x_n$ and $\mathbf{y} = y_1, ..., y_n$. In our case, $\mathbf{x}$ corresponds to the values annotated in our dataset for each dimension and $\mathbf{y}$ to those predicted by participant systems. The formula for *r* is as follows:

$$r(\mathbf{x}, \mathbf{y}) := \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

where $\bar{x}$ and $\bar{y}$ are respectively the mean value of $\mathbf{x}$ and $\mathbf{y}$.

MAE is a measure of errors between a couple of observations describing the same phenomenon (in this case the annotated values of a certain emotional dimension in the dataset, and those predicted). The formula for *MAE* is as follows:

$$MAE(\mathbf{x}, \mathbf{y}) := \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i| \quad (2)$$

The baselines for both sub-tasks have been built fine-tuning to a regression a BERT model available on HuggingFace[2], with a learning rate of 1e-05.

5. Results

We received submissions from two teams. Both of them participated to sub-task B, and only one to sub-task A. In total, 5 runs were submitted, constrained and unconstrained. In Table 5 we report the results for *r* and *MAE* in sub-task A, in Table 6 those relative to sub-task B, along with our baselines. We appended a suffix to distinguish the ID of the submitted run and another one to identify constrained ('_C_') and unconstrained ('_U_') runs.

Regarding sub-task A, the ISTC-CNR team obtained the best *r* score in the Valence dimension with his second run. Anyway, our baseline had better results in every other dimension and metric.

Concerning sub-task B, the team extremITA achieved the best results in all metrics and dimensions with their second run, with the exception of Arousal and Dominance's *r*, where our baseline performed slightly better.

6. Discussion

The teams of the EmotivITA challenge were invited to describe their solution in a technical report; in this section

[2]https://huggingface.co/dbmdz/bert-base-italian-xxl-uncased, last access 06-20-2023.

Table 5
Results of the submissions for sub-task A.

System	V *r*	A *r*	D *r*	V *MAE*	A *MAE*	D *MAE*
ISTC-CNR_1_C	0.800	0.593	0.609	0.393	0.432	0.399
ISTC-CNR_2_C	**0.809**	0.587	0.624	0.382	0.433	0.387
baseline	0.807	**0.643**	**0.643**	**0.313**	**0.321**	**0.285**

Table 6
Results of the submissions for sub-task B.

System	V *r*	A *r*	D *r*	V *MAE*	A *MAE*	D *MAE*
ISTC-CNR_1_C	0.800	0.594	0.623	0.400	0.421	0.367
extremITA_1_U	0.708	0.430	0.548	0.327	0.395	0.297
extremITA_2_U	**0.811**	0.633	0.630	**0.272**	**0.296**	**0.266**
baseline	**0.811**	**0.652**	**0.654**	0.859	0.859	0.859

we compare participant systems based on their architectures.

The ISTC-CNR team proposed a method based on Natural Language Inference (NLI). More specifically, they used a multilingual MNLI-XML-RoBERTa model grounded on XML-RoBERTa [35], which was fine-tuned on a version of the MNLI dataset [36] automatically translated to Italian. The model was adapted for the regression task replacing its last linear layer. During training, sentences from the EmoITA dataset were used as premises. Then, for sub-task A, three different models were conceived, with three different prompts acting as hypotheses for the NLI process and targeting the VAD dimensions. The prompt for Valence was “quanta positività esprime la frase?" (how much positivity does the sentence convey?), the one for Arousal “quanto è eccitante la frase?" (how exciting is the sentence?) and the one for Domination “quanto è controllata l’emozione" (how controlled is the emotion?). For sub-task B, a single model was used adopting the prompt “valence, arousal, dominance dell’emozione?" (valence, arousal, and dominance of the emotion?). The two runs submitted for sub-task A differ in that the first one only utilized 99% of the training set made available, while the second one utilized it entirely. As we can see in Table 5, the results were better with this last configuration. The only run submitted by the team for sub-task B exploited the entire training set. All runs were produced according to the constrained modality.

The extremITA team only participated to sub-task B, with two unconstrained runs. Both their models were trained on the union of all the datasets in the shared tasks at EVALITA 2023. The first one adopts an Encoder-Decoder architecture based on IT5 [37], a T5 model [38] pre-trained on Italian texts. The model was fine-tuned concatenating the name of the shared task as a prefix, followed by an input sentence from the EmoITA dataset. The output, in the case of the EmotivITA task, was constituted by the predicted VAD values. A similar approach was used for every task in EVALITA 2023. The second architecture is a Decoder that adopts instruction-tuning, based on a large language model, the LLaMA [39]. The model was trained using Low-Rank Adaptation [40] on Italian translations of the instructions originally developed for Alpaca [41], which also builds on LLaMA. It was then fine-tuned using instructions specific to the addressed EVALITA task. In the case of EmotivITA sub-task B, the sentence from the EmoITA dataset was paired with a prompt in the form of the instruction: “Scrivi quanta valenza è espressa in questo testo su una scala da 1 a 5, seguito da quanto stimolo è espresso in questo testo su una scala da 1 a 5, seguito da quanto controllo è espresso in questo testo in una scala da 1 a 5" (Rate how much valence is expressed in this text on a scale from 1 to 5, followed by how much arousal is expressed in this text on a scale from 1 to 5, followed by how much dominance is expressed in this text on a scale from 1 to 5). This second model obtained generally better performance than the first one as showcased in Table 6, but it also demanded 144 hours of training (on the entire EVALITA dataset), whereas the one based on IT5 only required 12 hours.

Quite interestingly the model proposed by the ISTC-CNR team and the second one proposed by the extremITA team both leverage prompting in natural language and no task-specific architectural designs (with the exception of the replacement of the last layer in the MNLI-XML-RoBERTa model), proving the efficacy of this approach. On the other hand, one could argue that the main limitations of the ISTC-CNR method was precisely the chosen prompts, as concepts like Valence, Arousal and Dominance are not easy to describe. When evaluating the extremITA proposal, instead, one could wonder about the sustainability of a 144 hours training process.

Anyway, we observe that the baselines obtained fine-tuning the BERT model were not outperformed by the proposed systems: maybe the upper limit for the regression problem with such a large dataset as EmoITA has

been reached, at least for the moment. It is also worth mentioning that the scores are in line with those of the study representing the state-of-the-art [42] for the original English dataset, EmoBank, that obtained values of 0.838, 0.573 and 0.536 for r in the three dimensions.

One last remark is due; neither team explored the possible relations between the three emotion dimensions, which was actually one of the purposes of sub-task B, and remains as a subject for future studies.

7. Conclusion

We presented the first shared task on Dimensional and Multidimensional Emotion Analysis for Italian and discussed the development of the first dedicated Italian dataset, EMoITA, based on the VAD model. EmoITA was obtained by manual translation and annotation of the EmoBank dataset, performed by Italian native speakers. The participating systems leveraged NLI, the Encoder-Decoder architecture and Large Language Models to address the regression problems, obtaining results that are similar to those of the state-of-the-art for the English counterpart of the dataset.

We hope that the proposal of our task and the availability of a new Italian dataset for EA will foster studies in this relevant field of NLP. In this spirit, the development and test set, as well as the complete dataset (licensed under CC-BY-SA 4.0), the script used for the baselines and for evaluation will be made available to the public soon; more details on EmotivITA can be found on the task website[3].

References

[1] R. W. Picard, Affective computing, MIT Press, Cambridge, Mass, 1997.

[2] M. Akhtar, D. Ghosal, A. Ekbal, P. Bhattacharyya, S. Kurohashi, All-in-one: Emotion, sentiment and intensity prediction using a multi-task ensemble framework, IEEE Transactions on Affective Computing 13 (2022) 285–297. doi:10.1109/TAFFC.2019.2926724.

[3] K. R. Scherer, Psychological Models of Emotion, in: J. C. Borod (Ed.), The Neuropsychology of Emotion, Oxford University Press, New York, 2000, pp. 137–162.

[4] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, SemEval-2014 task 4: Aspect based sentiment analysis, in: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Association for Computational Linguistics, Dublin, Ireland, 2014, pp. 27–35. URL: https://aclanthology.org/S14-2004. doi:10.3115/v1/S14-2004.

[5] S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, C. Cherry, SemEval-2016 task 6: Detecting stance in tweets, in: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), Association for Computational Linguistics, San Diego, California, 2016, pp. 31–41. URL: https://aclanthology.org/S16-1003. doi:10.18653/v1/S16-1003.

[6] A. Cignarella, M. Lai, C. Bosco, V. Patti, P. Rosso, SardiStance @ EVALITA2020: Overview of the Task on Stance Detection in Italian Tweets, in: V. Basile, D. Croce, M. Maro, L. C. Passaro (Eds.), EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020, Accademia University Press, 2020, pp. 177–186. URL: http://books.openedition.org/aaccademia/7084. doi:10.4000/books.aaccademia.7084.

[7] P. Basile, D. Croce, V. Basile, M. Polignano, Overview of the EVALITA 2018 Aspect-based Sentiment Analysis task (ABSITA), in: T. Caselli, N. Novielli, V. Patti, P. Rosso (Eds.), EVALITA Evaluation of NLP and Speech Tools for Italian, Accademia University Press, 2018, pp. 10–16. URL: http://books.openedition.org/aaccademia/4451. doi:10.4000/books.aaccademia.4451.

[8] P. Ekman, Basic Emotions, in: T. Dalgleish, M. J. Power (Eds.), Handbook of Cognition and Emotion, John Wiley & Sons, Ltd, Chichester, UK, 2005, pp. 45–60. URL: https://onlinelibrary.wiley.com/doi/10.1002/0470013494.ch3. doi:10.1002/0470013494.ch3.

[9] J. A. Russell, A. Mehrabian, Evidence for a three-factor theory of emotions, Journal of Research in Personality 11 (1977) 273–294. URL: https://linkinghub.elsevier.com/retrieve/pii/009265667790037X. doi:10.1016/0092-6566(77)90037-X.

[10] E. Cambria, A. Livingstone, A. Hussain, The hourglass of emotions, in: A. Esposito, A. M. Esposito, A. Vinciarelli, R. Hoffmann, V. C. Müller (Eds.), Cognitive Behavioural Systems, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 144–157.

[11] P. Kuppens, F. Tuerlinckx, J. A. Russell, L. F. Barrett, The relation between valence and arousal in subjective experience., Psychological Bulletin 139 (2013) 917–940. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/a0030811. doi:10.1037/a0030811.

[12] A. B. Warriner, V. Kuperman, M. Brysbaert, Norms of valence, arousal, and dominance for 13,915 English lemmas, Behavior Research Methods 45 (2013) 1191–1207. URL: http://link.springer.com/10.3758/s13428-012-0314-x. doi:10.

[3] Repository: https://github.com/GiovanniGafa/EmoITA. Website: https://sites.google.com/view/emotivita

3758/s13428-012-0314-x.

[13] S. Mohammad, F. Bravo-Marquez, M. Salameh, S. Kiritchenko, SemEval-2018 task 1: Affect in tweets, in: Proceedings of the 12th International Workshop on Semantic Evaluation, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 1–17. URL: https://aclanthology.org/S18-1001. doi:10.18653/v1/S18-1001.

[14] A. Chatterjee, K. N. Narahari, M. Joshi, P. Agrawal, SemEval-2019 task 3: EmoContext contextual emotion detection in text, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 39–48. URL: https://aclanthology.org/S19-2005. doi:10.18653/v1/S19-2005.

[15] O. Araque, L. Gatti, J. Staiano, M. Guerini, DepecheMood++: A Bilingual Emotion Lexicon Built Through Simple Yet Powerful Techniques, IEEE Transactions on Affective Computing 13 (2022) 496–507. URL: https://ieeexplore.ieee.org/document/8798675/. doi:10.1109/TAFFC.2019.2934444.

[16] M. Montefinese, E. Ambrosini, B. Fairfield, N. Mammarella, The adaptation of the Affective Norms for English Words (ANEW) for Italian, Behavior Research Methods 46 (2014) 887–903. URL: https://link.springer.com/10.3758/s13428-013-0405-3. doi:10.3758/s13428-013-0405-3.

[17] L. Passaro, L. Pollacci, A. Lenci, ItEM: A Vector Space Model to Bootstrap an Italian Emotive Lexicon, Second Italian Conference on Computational Linguistics CLiC-it 2015 II (2015).

[18] A. Bolioli, F. Salamino, V. Porzionato, Social Media Monitoring in Real Life with Blogmeter Platform, in: C. Battaglino, C. Bosco, E. Cambria, R. Damiano, V. Patti, P. Rosso (Eds.), Proceedings of the First International Workshop on Emotion and Sentiment in Social and Expressive Media: approaches and perspectives from AI (ESSEM 2013) A workshop of the XIII International Conference of the Italian Association for Artificial Intelligence (AI*IA 2013), Turin, Italy, December 3, 2013, volume 1096 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2013, pp. 156–163. URL: http://ceur-ws.org/Vol-1096/paper12.pdf.

[19] E. Borelli, D. Crepaldi, C. A. Porro, C. Cacciari, The psycholinguistic and affective structure of words conveying pain, PLOS ONE 13 (2018) e0199658. URL: https://dx.plos.org/10.1371/journal.pone.0199658. doi:10.1371/journal.pone.0199658.

[20] S. M. Mohammad, P. D. Turney, Crowdsourcing a word-emotion association lexicon, Computational Intelligence 29 (2013) 436–465.

[21] S. M. Mohammad, Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words, in: Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL), Melbourne, Australia, 2018.

[22] Celli, Fabio, Riccardi, Giuseppe, Ghosh, Aridam, CorEA: Italian news corpus with emotions and agreement, in: Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 and of the Fourth International Workshop EVALITA 2014 9-11 December 2014, Pisa, PISA UNIVERSITY PRESS, 2014. URL: http://clic2014.fileli.unipi.it/proceedings/Proceedings-CLICit-2014.pdf. doi:10.12871/CLICIT2014120.

[23] Z. Shibingfeng, F. Francesco, G. Federico, B.-C. Alberto, B. Paolo, P. Angelo, AriEmozione2.0, 2022. URL: https://zenodo.org/record/7097913. doi:10.5281/ZENODO.7097913.

[24] R. Sprugnoli, MultiEmotions-It: a New Dataset for Opinion Polarity and Emotion Analysis for Italian, in: J. Monti, F. dell'Orletta, F. Tamburini (Eds.), Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021, volume 2769 of *CEUR Workshop Proceedings*, CEUR-WS.org, Torino, 2020. URL: http://ceur-ws.org/Vol-2769/paper_08.pdf. doi:10.4000/books.aaccademia.8910.

[25] O. Araque, S. Frenda, D. Nozza, V. Patti, R. Sprugnoli, Emit at evalita2023: Overview of the categorical emotion detection in italian social media task, in: M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi (Eds.), Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[26] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[27] R. Plutchik, A General Psychoevolutionary Theory of Emotion, in: Theories of Emotion, Elsevier, 1980, pp. 3–33. URL: https://linkinghub.elsevier.com/retrieve/pii/B9780125587013500077. doi:10.1016/B978-0-12-558701-3.50007-7.

[28] R. Mukherjee, A. Naik, S. Poddar, S. Dasgupta, N. Ganguly, Understanding the role of affect dimensions in detecting emotions from tweets: A multi-task approach, CoRR abs/2105.03983 (2021). URL: https://arxiv.org/abs/2105.03983. arXiv:2105.03983.

[29] J. Wang, L.-C. Yu, K. R. Lai, X. Zhang, Dimensional sentiment analysis using a regional CNN-LSTM model, in: Proceedings of the 54th Annual Meet-

ing of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 225–230. URL: https://aclanthology.org/P16-2037. doi:10.18653/v1/P16-2037.

[30] S. Buechel, U. Hahn, EmoBank: Studying the Impact of Annotation Perspective and Representation Format on Dimensional Emotion Analysis, in: M. Lapata, P. Blunsom, A. Koller (Eds.), Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers, Association for Computational Linguistics, 2017, pp. 578–585. URL: http://aclweb.org/anthology/E17-2092. doi:10.18653/v1/E17-2092.

[31] N. Ide, C. Baker, C. Fellbaum, C. Fillmore, R. Passonneau, MASC: the manually annotated sub-corpus of American English, in: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), European Language Resources Association (ELRA), Marrakech, Morocco, 2008. URL: http://www.lrec-conf.org/proceedings/lrec2008/pdf/617_paper.pdf.

[32] C. Strapparava, R. Mihalcea, SemEval-2007 task 14: Affective text, in: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Association for Computational Linguistics, Prague, Czech Republic, 2007, pp. 70–74. URL: https://aclanthology.org/S07-1013.

[33] M. M. Bradley, P. J. Lang, Measuring emotion: The self-assessment manikin and the semantic differential, Journal of Behavior Therapy and Experimental Psychiatry 25 (1994) 49–59. URL: https://linkinghub.elsevier.com/retrieve/pii/0005791694900639. doi:10.1016/0005-7916(94)90063-9.

[34] P. J. Lang, Behavioral treatment and bio-behavioral assessment: Computer applications, in: J. B. Sidowski, J. H. Johnson, T. A. Williams (Eds.), Technology in mental health care delivery systems, Norwood, NJ: Ablex Publishing, 1980, pp. 119–137.

[35] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, CoRR abs/1911.02116 (2019). URL: http://arxiv.org/abs/1911.02116. arXiv:1911.02116.

[36] A. Williams, N. Nangia, S. Bowman, A broad-coverage challenge corpus for sentence understanding through inference, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 1112–1122. URL: https://aclanthology.org/N18-1101. doi:10.18653/v1/N18-1101.

[37] G. Sarti, M. Nissim, IT5: Large-scale text-to-text pretraining for italian language understanding and generation, ArXiv preprint 2203.03759 (2022). URL: https://arxiv.org/abs/2203.03759.

[38] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, 2020. arXiv:1910.10683.

[39] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. arXiv:2302.13971.

[40] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, 2021. arXiv:2106.09685.

[41] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T. B. Hashimoto, Stanford alpaca: An instruction-following llama model, https://github.com/tatsu-lab/stanford_alpaca, 2023.

[42] S. Park, J. Kim, S. Ye, J. Jeon, H. Y. Park, A. Oh, Dimensional emotion detection from categorical emotion, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 4367–4380. URL: https://aclanthology.org/2021.emnlp-main.358. doi:10.18653/v1/2021.emnlp-main.358.

ISTC-CNR at EmotivITA: Towards Better Dimensional and multi-dimensional Analysis of VAD Emotions*

Luana Bulla[1,2,*,†], Misael Mongiovì[2,†]

[1]University of Catania, Italy
[2]National Research Council, Italy

Abstract

Emotions are an integral part of human communication, and accurately detecting and interpreting them from textual data holds significant potential for numerous applications. The Valence-Arousal-Dominance (VAD) dimensional model provides a rich framework for capturing the nuanced emotional states conveyed through text. This paper presents an in-depth exploration of text-based emotion detection using VAD analysis and machine learning techniques. In this paper, we propose a novel machine-learning model specifically designed to detect VAD dimensions from textual data. Through empirical evaluation and comparison with existing methods, we analyze the results of our proposed model in accurately identifying emotions expressed in text. This paper describes the ISTC-CNR participation in the EmotivITA task at EVALITA 2023, showcasing our team's efforts and findings in the context of the competition.

Keywords

NLP, Machine Learning, NLI, Emotion Regression

1. Introduction

Emotions constitute a complex facet of human experience, influencing our perception, behavior, and overall well-being. Traditionally, emotions have been studied using categorical models [1], which assign discrete labels to distinct emotional states. However, the limitations of such models in capturing the richness and variability of emotions led to the development of dimensional models. Among these, the VAD dimensional model [2] has gained considerable attention due to its ability to provide a multi-dimensional representation of emotions. According to the VAD paradigm, emotions may be classified along three main dimensions: valence, which describes the pleasantness or unpleasantness of an emotional experience; arousal, which reflects the level of physiological activation associated with emotion; and dominance, which denotes the degree of control or influence exerted by an emotion. This dimensional framework offers a more fine-grained understanding of emotional experiences, enabling researchers to capture the subtle nuances and interplay of emotions.

The primary objective of this paper is to explore text-based emotion detection by leveraging the VAD dimensional model through the fine-tuning of a multi-lingual MNLI-XML-RoBERTa model. The MNLI-XML-RoBERTa model is based on the well-known architecture of XML-RoBERTa [3] and has been fine-tuned on an Italian version of the MNLI dataset [4] for solving a Natural Language Inference (NLI) task. In this study, we investigate the model's inference capability and its adaptability to handle multiple languages to address the VAD regression task in both single- and multi-dimensional scenarios. To achieve our goal, we apply the model to the Italian version of the EmoBank dataset (EmoITA) [5], which consists of 10,000 short text documents tagged with VAD emotional scores. We propose a specific configuration of the MNLI-XML-RoBERTa model for both mono- and multi-dimensional scenarios. In the first scenario, we train three separate models, each capable of predicting one of the Valence, Arousal, or Dominance values using only the target dimension during training. For multi-dimensional emotional analysis, we develop a single model that predicts all three values jointly by utilizing training data from all three dimensions.

We present our model at EmotivITA[1], a task in the EVALITA 2023 campaign [6], validating and discussing our results while comparing them with the baseline provided by the organizers for both sub-task A (Dimensional emotion regression) and B (Multi-dimensional emotion regression)[2].

The paper is organized as follows: Section 2 introduces our novel machine learning model for VAD-based emotion detection, providing an insightful description of the model architecture. Section 3 presents the experimental setup, evaluation metrics, and results, comparing the

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
*Corresponding author.
†These authors contributed equally.
luana.bulla@phd.unict.it (L. Bulla); misael.mongiovi@cnr.it (M. Mongiovì)
0000-0003-1165-853X (L. Bulla); 0000-0003-0528-5490 (M. Mongiovì)
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

[1]https://sites.google.com/view/emotivita
[2]https://sites.google.com/view/emotivita/the-tasks

CEUR-WS.org/Vol-3473/paper6.pdf

performance of the VAD MNLI-XML-ROBERTA-based system with the baseline BERT-based model provided by the organizers. Finally, Section 4 summarizes the key findings, discusses the practical implications, and outlines potential directions for future research.

2. Description of the Systems

We describe our NLI-based method for recognizing fine-grained emotions along the continuous VAD dimensions. Our approach encompasses both the dimensional (sub-task A) and multi-dimensional (sub-task B) VAD regression tasks, following a unified methodology framework. To accomplish this, we introduce a supervised model developed within a constrained mode, leveraging the EmoITA. Section 2.1 provides a comprehensive overview of our methodology, offering detailed insights into the key components and techniques employed in our NLI-based approach. Section 2.2 provide an overview of the prompts utilized during the training phase.

2.1. VAD MNLI-XML-RoBERTa model

We present the VAD MNLI-XML-RoBERTa model, a fine-tuned version of the multilingual XML-RoBERTa system, which has been developed using an automatically translated Italian version of the MNLI corpus[3]. To adapt the model for the regression task, we replace the last linear layer of the MNLI-XML-RoBERTa model. During training, the parameters of the last linear layer are initialized from scratch and learned, while the remaining parameters are fine-tuned. To tune the hyperparameter, we train the model on a portion of the training set provided by the organizers in a preliminary phase. Our model utilizes input sentences as premises, accompanied by predefined fixed prompts that serve as hypotheses. These prompts aim to guide the model's training process by focusing on specific assignments. The VAD MNLI-XML-RoBERTa model is trained with a learning rate of 1e-5, a batch size of 64, and AdamW as the optimizer. The Mean Squared Error (MSE) is adopted as the loss function, measuring the disparity between the model's predictions and the actual values[4].

2.2. Prompting

We introduce a set of tailored prompts designed to predict the values of valence, arousal, and dominance in text. For the dimensional emotion analysis task, we develop three prompts, each targeting a specific emotion dimension. The prompt for valence is "quanta positività esprime la frase?" (how much positivity does the sentence convey?). For arousal, the prompt is "quanto è eccitante la frase?" (how exciting is the sentence?). Lastly, for dominance, the prompt is "quanto è controllata l'emozione?" (how controlled is the emotion?). These prompts are crafted to succinctly encapsulate the essence of each emotional dimension using easily understood language. Their purpose is to guide the model's training process toward discerning specific emotional dimensions accurately. In the multi-dimensional emotion analysis task, we utilize a single prompt to predict all emotion dimensions. The prompt used is "valence, arousal, dominance dell'emozione?" (valence, arousal, dominance of the emotion?). In this case, the prompt assists the model in comprehending the comprehensive framework of the task during the training phase, providing contextual information. By leveraging these prompts, our objective is to enhance the model's capacity to capture and comprehend the intricate nuances of emotional dimensions in text.

3. Results and Evaluation

We conduct an experimental analysis to assess the effectiveness of our proposed approach and compare it with the baseline method provided by the competition's organizers using EmoITA [5]. In Section 3.1, we present the results of the dimensional emotion analysis, focusing on the individual evaluation of valence, arousal, and dominance. Here, we examine the performance of our approach and provide a comprehensive analysis of its efficacy. Moving on to Section 3.2, we shift our focus to multi-dimensional emotion analysis. In this section, we evaluate the performance of our approach in capturing and predicting emotional dimensions collectively.

3.1. Dimensional Emotion Analysis

We train three separate VAD MNLI-XML-RoBERTa models, each focused on detecting the degree of valence, arousal, and dominance values individually. During training, each model focuses on a specific target dimension and is associated with a specific prompt (Sect. 2).

Tables 3.1 and 3.1 showcase the results obtained by evaluating our MNLI-based model on 2′063 items from the Italian EmoBank test set provided by the organizers. For the dimensional emotion analysis task, we submitted two different runs corresponding to distinct configurations of the fine-tuned VAD MNLI-XML-RoBERTa model. During the model training phase, the first run incorporates 99% of the data from the development set, while the second run utilizes the entire development set. Each emotional dimension is evaluated based on Pearson's correlation coefficient (Table 3.1) and Mean Absolute Errors (Table 3.1). To assess the effectiveness of our model, we

[3] https://huggingface.co/Jiva/xlm-roberta-large-it-mnli
[4] https://pytorch.org/docs/stable/generated/torch.nn.MSELoss.html

Table 1
Pearson's r correlation score for valence, arousal and dominance values for the baseline BERT-based model and the MNLI-based systems in the Dimensional emotional detection task.

Model	Valence	Arousal	Dominance
Baseline	0.807	**0.643**	**0.643**
VAD MNLI-XML-RoBERTa (run 1)	0.800	0.593	0.609
VAD MNLI-XML-RoBERTa (run 2)	**0.809**	0.587	0.624

Table 2
Mean Absolute Error (MAE) score for valence, arousal and dominance values for the baseline BERT-based model and the MNLI-based systems in the Dimensional emotional detection task.

Model	Valence	Arousal	Dominance
Baseline	0.313	0.321	0.285
VAD MNLI-XML-RoBERTa (run 1)	0.393	0.432	0.399
VAD MNLI-XML-RoBERTa (run 2)	0.382	0.433	0.387

compare its performance against a baseline BERT-based model provided by the organizers. Our NLI-based model demonstrates slightly higher Pearson's correlation values for valence compared to the baseline, while the other metrics show slightly lower scores. Notably, valence exhibits the highest performance in terms of correlation, while arousal and dominance present more challenges, with the baseline model achieving a correlation of only 0.64%.

3.2. Multi-dimensional Emotion Analysis

We present the performance of models designed for detecting valence, arousal, and dominance in a multi-dimensional setting. In this setting, both the baseline and MNLI-based models predict all three values by utilizing all three VAD dimensions during the training phase. For this task, we submitted a single run where we employed all the available development set data for training the model. We report the performances of the VAD MNLI-XML-RoBERTa system for each emotion dimension, evaluating them in terms of Pearson's correlation coefficient (Table 3.2) and Mean Absolute Error (Table 3.2). We compare our results with the baseline performance provided by the organizers.

Both the baseline and NLI-based models demonstrate superior performance in detecting valence, achieving a correlation of approximately 80%. As previously mentioned, the scores for the arousal and dominance dimensions are comparatively lower than valence. Specifically, the highest correlation values observed for arousal and dominance values are 0.652 and 0.654, respectively.

4. Discussion

As anticipated, performance varied significantly across different emotion values in both dimensional (i.e. sub-task A) and multi-dimensional (i.e. sub-task B) emotion regression tasks. The concept of valence demonstrated consistent interpretation, resulting in higher correlation scores for all models in both analyses. Conversely, detecting concepts like "arousal" and "dominance" proved more challenging in the context of short text documents. Comparing the two models, the BERT-based model yielded the best results, while the NLI-based model exhibited slightly lower performance in both tasks. This discrepancy can be attributed to the multilingual nature of the MNLI-XML-RoBERTa model, which grants it greater adaptability but limits its ability to identify nuanced and ambiguous terms like "dominance" and "excitement" within text. Additionally, the NLI-based models employed a generic prompt for every item in the dataset as a hypothesis. This approach potentially underutilized the model's capabilities as the prompt lacked the necessary explanatory power to effectively tackle challenging regression tasks. In conclusion, while the models demonstrated strengths in capturing valence, improvements are needed to enhance their performance in discerning arousal and dominance. This includes refining the model's adaptability to diverse linguistic contexts and developing more informative prompts for comprehensive analysis of emotional dimensions.

5. Conclusion

In our study, we propose a novel approach for detecting valence, arousal, and dominance values in natural language short text using an MNLI-XML-RoBERTa-based model. Our method encompasses both dimensional and

Table 3
Pearson's r correlation score for valence, arousal and dominance values for the baseline BERT-based model and the MNLI-based systems in the multi-dimensional emotional detection task.

Model	Valence	Arousal	Dominance
Baseline	0.811	**0.652**	**0.654**
VAD MNLI-XML-RoBERTa	0.800	0.594	0.623

Table 4
Mean Absolute Error (MAE) score for valence, arousal and dominance values for the baseline BERT-based model and the MNLI-based systems in the multi-dimensional emotional detection task.

Model	Valence	Arousal	Dominance
Baseline	0.859	0.859	0.859
VAD MNLI-XML-RoBERTa	0.400	0.421	0.367

multi-dimensional regression scenarios, leveraging a fine-tuned XML-RoBERTa classifier trained and evaluated on the EmoITA. We present the results of both the dimensional and multi-dimensional configurations' models and compare them with the performance of the BERT-based model provided by the competition's organizers. This evaluation provides valuable insights into the effectiveness of our proposed approach in capturing and analyzing emotional dimensions in textual data.

References

[1] F. A. Acheampong, C. Wenyu, H. Nunoo-Mensah, Text-based emotion detection: Advances, challenges, and opportunities, Engineering Reports 2 (2020) e12189.

[2] J. A. Russell, A. Mehrabian, Evidence for a three-factor theory of emotions, Journal of research in Personality 11 (1977) 273–294.

[3] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, CoRR abs/1911.02116 (2019). URL: http://arxiv.org/abs/1911.02116. arXiv:1911.02116.

[4] A. Williams, N. Nangia, S. Bowman, A broad-coverage challenge corpus for sentence understanding through inference, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 1112–1122. URL: https://aclanthology.org/N18-1101. doi:10.18653/v1/N18-1101.

[5] G. Gafà, F. Cutugno, M. Venuti, EmotivITA at EVALITA2023: Overview of the Dimensional and Multidimensional Emotion Analysis Task, in: M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi (Eds.), Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[6] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

PoliticIT at EVALITA 2023: Overview of the Political Ideology Detection in Italian Texts Task

Daniel Russo[1,2], Salud María Jiménez-Zafra[3], José Antonio García-Díaz[4], Tommaso Caselli[5], Marco Guerini[1], L. Alfonso Ureña-López[3] and Rafael Valencia-García[4]

[1]LanD, Fondazione Bruno Kessler, Via Sommarive 18, Povo, Trento, Italy
[2]University of Trento, Italy
[3]SINAI, Universidad de Jaén, Spain
[4]UMUTeam, Universidad de Murcia, Spain
[5]CLCG, University of Groningen, Netherlands

Abstract
This paper presents the PoliticIT 2023 shared task, organised at EVALITA 2023 workshop. The task aims to extract politicians' ideology information from a set of tweets in Italian framed as a binary and a multiclass classification. The task is designed to be privacy-preserving and it is accompanied by a subtask targeting the identification of self-assigned gender as a demographic trait. The PoliticIT task attracted 7 teams that registered for the task, submitted results and presented working notes describing their systems. Most of the teams proposed transformer-based approaches, while some of them also used traditional machine learning algorithms or even a combination of both.

Keywords
Author profiling, Political ideology, Author analysis, Demographic and psychographic traits

1. Introduction and Motivations

The study of the political discourse on Social Media Platforms is of paramount importance in order to understand where society is heading. Political discourse is by definition ideologically based and political ideologies are spread with discourse [1]: for this reason, the analysis of the latter cannot go without the understanding of the former.

Political ideology is defined as a psychographic trait that can help comprehend both individual and social behaviour, including moral and ethical values, attitudes, biases, and prejudices [2]. In fact, this trait helps understanding how individuals think that society should be organised and has a strong relationship with personality traits as demonstrated in [3]. For instance, they found that conscientiousness is strongly correlated with the right wing, whereas openness to experience and agreeability were notably more correlated to the left wing. Moreover, political ideology has a great influence in the daily lives of each citizen. For example, [4] found a correlation between political ideology and the attitude of citizens to vaccination campaigns. Still, citizens react to the political messages they are exposed to. Therefore studying how politicians spread their ideology using social media discourses is useful to better analyse the policies and perspectives that are proposed on how society should be organized and work.

In this scenario, the PoliticIT shared task organized at EVALITA 2023 [5] aims to extract political ideology information from texts. To this end, an author profiling task is proposed. In particular, the PoliticIT task focuses on the identification of the political leaning and the self-assigned gender of the author from a binary and multiclass perspective.

In recent years, several shared tasks for author analysis have been organized under the PAN workshop series [6], targeting author attribution, bot detection, gender detection, and author obfuscation, among others. Other initiatives, such as the PoliticES shared task [7], have focused on capturing other traits such as the political ideology expressed in a message. PoliticIT is a twin-task of PoliticES and aims at analysing political ideology while being privacy-preserving. For this reason, a novel methodology of text clustering, that creates "virtual users" has been used on top of traditional anonymisation procedures.

The rest of the paper is organized as follows. Section

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
drusso@fbk.eu (D. Russo); sjzafra@ujaen.es (S. M. Jiménez-Zafra); joseantonio.garcia8@um.es (J. A. García-Díaz); t.caselli@rug.nl (T. Caselli); guerini@fbk.eu (M. Guerini); laurena@ujaen.es (L. A. Ureña-López); valencia@um.es (R. Valencia-García)
https://www.rug.nl/staff/t.caselli (T. Caselli); https://www.marcoguerini.eu (M. Guerini)
0009-0006-9123-5316 (D. Russo); 0000-0003-3274-8825 (S. M. Jiménez-Zafra); 0000-0002-3651-2660 (J. A. García-Díaz); 0000-0003-2936-0256 (T. Caselli); 0000-0003-1582-6617 (M. Guerini); 0000-0001-7540-4059 (L. A. Ureña-López); 0000-0003-2457-1791 (R. Valencia-García)
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper7.pdf

2 describes the PoliticIT shared task. Section 3 presents the dataset provided in the competition. Section 4 summarises the participant approaches. Section 5 shows the results and a discussion thereof. Finally, Section 6 concludes the paper with a discussion of the most interesting outcomes of this task and possible future works.

2. Task Description

The PoliticIT task is structured along three subtasks:

- **Subtask A** - *Self-assigned Gender*: Given a message, the system must assign a value for the gender of the author. The set of labels has been determined according to the personal web pages of the politicians of the Italian Parliament. The task has been framed as a binary classification task with **M** for men and **F** for women.
- **Subtask B** - *Political Ideology (binary)*: systems are required to determine the political orientation of a message; the binary version of the task presents two macro-categories: **Left** and **Right**.
- **Subtask C** - *Political Ideology (multi-class)*: this subtask presents a more fine-grained set of labels for the political orientation expressed by a given message. In this case, we employed four labels: **Left**, **Moderate-Left**, **Moderate-Right**, and **Right**.

PoliticIT was organized through CodaLab.[1] The run of the task is divided into three phases: (i) Practice, (ii) Evaluation, and (iii) Post-evaluation. In the Practice phase, the participants were initially provided with a subset of the training data in order to familiarise themselves with the training data format. During this phase, we also provided a notebook comprising the code for our Logistic Regression baseline, as a starting point for the development of more efficient systems. The full training set was released in February 2023. Currently, the task is in its post-evaluation phase, where participation is publicly open to other teams and research groups from the community.

3. Datasets and Format

This section provides the reader with an overview of the dataset proposed for the PoliticIT 2023 shared task along with a comprehensive description of the modalities employed for creating it.

[1] https://codalab.lisn.upsaclay.fr/competitions/8507

3.1. Data Collection

The dataset was collected from the Twitter timelines[2] of Italian politicians using the UMUCorpusClassifier [8], following a strategy similar to the one adopted in PoliticES 2022 [7] and in [9]. In particular, the data refer to the politicians from the legislature XIX of the Italian Republic. The list of deputies, senators, and ministers was taken from the institutional websites of the Italian Parliament[3] and Government.[4] All the politicians' Twitter accounts were manually retrieved, as they are not reported on the institutional websites. We discarded politicians that did not have a Twitter account or that were highly inactive on this social media (i.e. whose accounts present very few or old tweets). The time window for the corpus compilation was December 2022 as the oldest date, but no start date was set. In the first iteration, we compiled 371,822 tweets from 468 politicians between November 2010 and December 2022. The average number of tweets per politician is 794.49 but with a large standard deviation of 847.12, which suggest that not all politicians are equally active on Twitter. Thus, we decided to remove from the dataset those politicians with less than 25 tweets, leaving a total of 408 politicians.

To balance the number of tweets per politician, we first removed those tweets that are not written in Italian. To detect the language, we employed FastText language identification model [10]. Secondly, we removed the documents that shared content from news websites without retweeting. To do this, we discarded tweets that contained mentions of news websites, by detecting linguistic clues within the text, such as the pipe symbol, which is commonly employed by news websites to categorise their content. Thirdly, we selected tweets based on topics. An initial list of topics was extracted with BERTopic [11], a topic modelling technique for the creation of interpretable clusters based on Transformers and c-TF-IDF. In particular, we leveraged the Italian BERT model from [12]. We obtained a list of topics organised into 21 categories. This list was manually checked to introduce additional keywords for categories such as *European Union*, *immigration*, *energy*, *feminism*, *sports*, *mafia* or *religion*. Next, we identified which topics appeared in each tweet and prioritised those tweets that contained at least one topic. We then selected the tweets according to their topic in order to avoid any possible bias in the dataset.

[2] https://developer.twitter.com/en/docs/twitter-api/v1/tweets/timelines/api-reference/get-statuses-home_timeline

[3] Chamber of Deputies: https://www.camera.it/leg19/28
Senate of the Republic: https://www.senato.it/leg/19/BGT/Schede/Attsen/Sena.html

[4] https://www.governo.it/it/ministri-e-sottosegretari

3.2. Data Annotation and Anonymization

We enriched the dataset by assigning to each politician a label indicating their political ideology. Political ideologies have been directly derived from the politicians' affiliation party. In particular, the mapping from the politician to the political ideology was obtained through a two-step procedure:

1. Automatic labelling of politicians with their current political party affiliation. The party affiliation has been inferred from the parliamentary group to which the parliament party belongs. The data were extracted from the Italian institutional websites on October 31, 2023, thus they do not reflect changes in parliamentary groups following this date.
2. Mapping of the political parties to specific political ideology labels. The set of labels has been identified using Wikipedia.[5] We used four political ideology labels, i.e. *left*, *moderate left*, *right*, *moderate right*. Parties that are mapped in the centre, or cross-party, were nevertheless assigned one of the four aforementioned labels on the basis of their political alliances and the programme they presented during the 2022 Italian election campaign. The decision to "force" this classification was made to avoid excessive imbalance within each class. Therefore, we labelled "Movimento 5 Stelle" as *left*, whereas "Azione" and "Italia Viva" as *moderate left*.

Gender labels were assigned through three different approaches, depending on the source of the data. for the Italian deputies, gender was directly extracted from the institutional website, which allows the filtering of members according to this trait. The website of the Senate of the Republic does not clearly states the gender of the members. In this case, employed linguistic cues present on the personal page of each senator to infer the gender. Specifically, we looked at the Italian verb "nascere" in its past participle form as "nato" for the male label and "nata" for the female label. Finally, for the ministers, gender was manually assigned as the official Government website does not comprise this information and does not present helpful linguistic cues. In this case, ministers were labelled according to their biological sex.

Subsequently, to build a privacy-compliant approach we took a two-step procedure including *anonymization* and *clustering*:

- **Anonymized references in text** - References to politicians within Twitter mentions were anonymized by replacing them with the token `@user`. The rest of the in-text mentions were also replaced with the `@user` token. Consequently, the text traits cannot be guessed trivially by reading a politician's name and searching for personal information on the Internet. We also replaced the name of the political parties and of their Twitter accounts with the `POLITICAL_PARTY` token.
- **Clustering procedure** - Subsequently, we created clusters of texts by mixing some of the extracted tweets in order to prevent ethical and privacy issues related to author profiling on Twitter. All the clusters are composed of tweets written by different politicians that share the same traits under evaluation, i.e. political ideology and self-assigned gender. For this, we divide the politicians into training and test in order to prevent that tweets from the same politician from appearing both in training and validation. To generate a cluster, we first set their demographic and psychographic traits, and then randomly pick tweets from users that share these traits. Thereby, each cluster represents "virtual users", with their self-assigned gender (male, female) and political spectrum. For the latter, we labelled the data according to two axes: binary (left, right) and multiclass (left, moderate left, moderate right and right). At the end of this process, we obtained 1751 clusters with 80 tweets per cluster. It should be noted that the clusters from the training and test sets are independent to prevent machine learning approaches from identifying the authors rather than the demographic and psychographic traits.

3.3. Data Formats

The training and test sets are produced in a ratio of nearly 75%-25%. Table 1 presents a summary of the distribution of labels per subtask. In no case, the labels are evenly distributed. Male politicians are almost double the number of female politicians, and more than 200 politicians are from the left wing than from the right. As regards the multi-class ideology, moderate left and right are the most represented labels.

Ultimately, each entry of the PoliticIT dataset comprises four elements: a cluster id, the self-assigned gender label and the political ideology labels for binary and multiclass classification. The dataset is organised at tweet level. This means that each row represents one tweet. Each line also contains a `cluster id` to identify the cluster to which the tweet belongs, as well as the demographic and psychographic traits of the cluster. Examples are provided in Table 2. The full dataset, including the gold labels, is available on Codalab [6].

[5] https://it.wikipedia.org/wiki/Partiti_politici_italiani

[6] https://codalab.lisn.upsaclay.fr/competitions/8507

Table 1
PoliticIT: Data distribution per subclass for Train and Test splits.

Subtask	Label	Train	Test	Total
Subtask A - Gender	Male	810	318	1128
	Female	488	135	623
Subtask B - Ideology (binary)	Left	720	248	968
	Right	578	205	783
Subtask C - Ideology (multi-class)	Left	162	100	262
	Moderate left	558	148	706
	Moderate right	447	154	601
	Right	131	51	182

Table 2
Examples from our dataset. Each dataset entry comprises: the cluster id, the self-assigned gender label, and the political ideology labels, both binary (pib) and multiclass (pim).

cluster id	gender	pib	pim	tweet
72b...	female	left	left	Non possiamo permettere che gli imprenditori si ritrovino dal giorno alla notte non più competitivi e che le imprese vengano lasciate sole. E per questo come @user stiamo lavorando senza sosta.
8f5...	male	right	right	Se spacci, vai in carcere e ci resti!. Con [POLITICIAN] proposta #[POLITICAL_PARTY], presentata in Parlamento, città più sicure e più rispetto per le Forze dell'Ordine, che rischiano per arrestare i pusher e poi se li ritrovano per strada il giorno dopo.

4. Systems Overview

A total of seven teams participated in the PoliticIT task, with all teams involved in each of the subtasks. The majority of the participants represented academic institutions. An overview of the system approaches can be found in Table 3, while Section 4.1 provides further details on the systems proposed.

4.1. System Architectures

ExtremITA [13]. The team proposed two systems. The first system is based on Camoscio [14], the Italian version of the Standford Alpaca model [15], pre-trained to generate text as a response to users' instructions passed as input. The team performed further fine-tuning on triples <`task`, `input`, `output`>. More precisely, the model used the phrasal forms derived from the training data of all EVALITA 2023 challenges: the `task` is a linguistic description of the task to be solved, whereas the `input-output` pairs are task-specific. On the other hand, the second system is based on IT5 Transformer [16], for which fine-tuning was done on `input-output` pairs. More precisely, the model used the phrasal forms derived from the training data of all EVALITA 2023 challenges, where the `input-output` pair is task-specific.

INFOTEC-LaBD [17]. The team employed SVM classifiers with linear and nonlinear kernels. Specific attention was given to the data representation. Indeed, the authors employed low-dimensional projections to concisely represent the dataset and its associated labels. These vectors were used for training an SVM classifier.

INGEOTEC. The team implemented different configurations of Bag of Words (BoW) classifier in all the subtasks. In particular, for gender and political ideology multiclass classification, INGEOTEC employed a stack generalization approach leveraging three BoW classifiers: two BoW classifiers pre-trained on 5M Italian tweets, and a BoW classifier trained on PoliticIT training set. Instead, for the political ideology binary subtask, the team proposed a BoW classifier trained on the training set.

Teeeech. The team proposed three Transformer-based classifiers trained independently of each other. The authors did not specify which Transformer models were used.

Tübingen [18]. The team proposed two main approaches: an SVM-based approach and a Transformer-based approach. The former was the best-performing, i.e.,.e simple linear SVMs with sparse word/character n-gram features, trained separately for each task, only

Table 3
Overview of the participants' systems. For each team, we summarize the pre-trained language models employed, and the method(s) used for their runs.

	Language Model					Method					
Team	**dbmdz BERT Italian**	**AlBERTO**	**IT5**	**Camoscio**	**XLM-RoBERTA**	**SVM**	**BoW classifier**	**Fine-tuning**	**Low-dimensional projection**	**Prompting**	**Data augmentation**
extremITA			✓	✓						✓	
INFOTEC-LaBD						✓			✓		
INGEOTEC							✓				
Teeeech								✓			
Tübingen	✓				✓	✓		✓			✓
UMUTeam	✓							✓			
URJC_TEAM	✓	✓						✓			

using the data provided for the shared task. For the Transformer-based approach, the team experimented with Italian BERT monolingual model [12] and XLM-ROBERTA multilingual model [19] for the classification task, multi-task/multi-label learning, and a number of transfer learning approaches. As additional data they employed the Italian section of the ParlaMint corpus [20] and PoliticES data [7].

UMUTeam [21]. The team proposed a system based on the Italian BERT [12] fine-tuned on PoliticIT data. Demographic and psychographic traits were classified according to two strategies: (i) *mode*, which selects the most repeated label after classifying each user's tweet individually, and (2) *highest probability*, which selects the label with the highest probability.

URJC_TEAM [22]. The system developed by the team leverages three BERT models. Each model was pre-trained on different data according to the subtask it was employed. In particular, the model for gender classification model was trained using text data sourced from Wikipedia dump and the OPUS corpora collection [23]. The binary classification model used to identify political polarity was trained both on the OPUS corpora [24] and on the OSCAR corpus. Lastly, for the multilabel classification subtask, the AlBERTo language model [25] was employed.

5. Results and Discussion

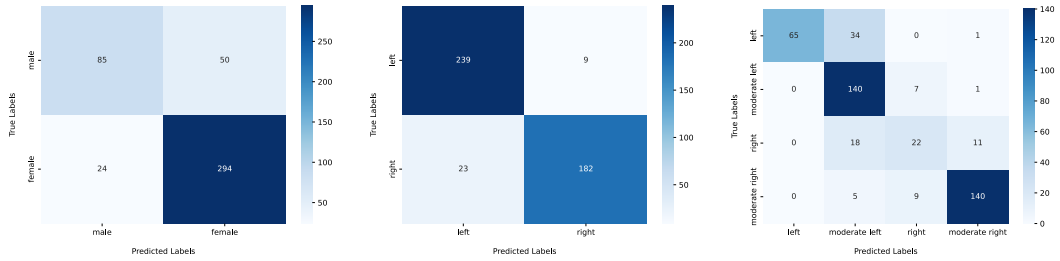
For each subtask, participants were allowed to make a maximum of 10 submissions in CodaLab during the Evaluation phase. For the ranking, participants were asked to select their best model. Systems were evaluated using standard evaluation metrics for text classification, namely Precision, Recall, and F1-score per class. The ranking is based on the arithmetic mean of the macro F1-scores of each subtask.

For all subtasks, we implemented the same baseline using TF–IDF based on word-unigrams, 50,000 as maximum features, minimum feature ratio of 10%, and applying lowercase. These features were trained using a Logistic Regression classifier. Table 4 summarises the leaderboard results and rankings.

Overall, all systems, with the sole exclusion of **URCJ_TEAM** on binary ideology classification, managed to outperform the baseline approach on all subtasks. We can notice that differences in performances across the top five systems are relatively large, ranging from a minimum of 0.024 points to a maximum of 0.071. This highlights a different trend when compared to other shared tasks where usually the top systems are all very close. On the other hand, a similar behaviour can be observed in PoliticES [7], where the range across the top five systems is 0.0.11 and 0.0827, suggesting that the task could present inherent challenges, especially for Subtask A (gender detection) and Subtask C (fine-grained political ideology classification).

Table 4
PoliticIT official leaderboard.

Team	Macro F1 (average)	F1-gender	F1-ideology (binary)	F1-ideology (multiclass)	Rank
Tübingen	0.824	0.792	0.928	0.751	1
INFOTEC-LaBD	0.800	0.824	0.860	0.717	2
extremITA	0.771	0.769	0.925	0.621	3
INGEOTEC	0.762	0.732	0.848	0.705	4
Teeeech	0.751	0.752	0.887	0.613	5
UMUTeam	0.704	0.712	0.866	0.533	6
URJC_TEAM	0.684	0.661	0.770	0.621	7
Baseline	0.569	0.536	0.809	0.362	–



(a) Gender (b) Ideology (binary) (c) Ideology (multiclass)

Figure 1: Confusion matrices for each subtask – self-assigned gender detection (1a), binary political ideology (1b) and multi-class political ideology classification (1c) – from the best performing system (team **Tübingen**).

Focusing on each subtask, it appears evident that binary classification of political ideologies is relatively easier when compared to the other two subtasks, with the best results being 0.928 obtained by **Tübingen**. The more fine-grained the distinction of political ideologies is the more challenging the task. This is not just an effect of having multiple classes and the distribution of the data, but it involves also the subtleties and nuances in policies across the four groups. In Figure 1 we display the confusion matrices obtained by the Tübingen team for the three classification subtasks. Focusing on the errors of Subtask C, multi-class ideology classification, we can notice that most of the errors concern misclassification of the "extremes" (i.e., **Left** and **Right**) into the **Moderate Left** category) rather than of the moderate positions (**Moderate Left** *vs.* **Moderate Right**). This indicates that - at least in their communication on Twitter - differences in the moderate political areas are stronger. A further noticeable result is the fact that messages from politicians on the **Right** spectrum tend to be assimilated mostly with the **Moderate Left** and **Moderate Right**, suggesting that the narratives of moderate groups tend to assimilate issues and expressions of the **Right**.

The identification of the gender of the author from a tweet is also quite challenging even if it is framed as a binary task. In this case, as illustrated by Figure 1a most of the errors affect the classification of male politicians into females. The best system, **INFOTEC-LaBD**, obtains 0.824 macro F1, with a positive Δ from the second best system (**Tübingen**) of 0.032 points.

6. Conclusions

In this paper, we have summarized the outcomes of the first edition of the PoliticIT task at EVALITA 2023. PoliticIT targets the identification of the political ideology and gender of the author of a tweet. Political ideology is a psychographic trait that can be used to understand individual and social behaviour, and thus contribute to a better understanding of the society. The task introduces an innovative method concerning the anonymization of users to preserve privacy, allowing the investigation of these sensitive topics in a fair and ethical way.

PoliticIT has seen the participation of seven teams, five of whom submitted a full report describing their approach. The results indicate that fine-grained political ideology distinction is more challenging than binary classification between two extreme values. This appears to be due to the absorption of the narratives of the extremes into the moderate positions, especially from the **Moderate Left**. This is a datum that could provide political scientists and sociologists additional insights into the

evolution of the Italian political systems. A further aspect that raises interest concerns the errors in Subtask A, gender classification. In this case, most of the errors are male politicians classified as females. A deep exploration of the communication styles of the two genres and their correlation with political affiliations is a promising path to better understanding this behaviour.

As expected, approaches based on Transformers are the trend solutions presented by participating teams, but some of them also used feature-based linear machine learning systems. It is quite impressive that the best performing team, **Tübingen**, uses linear SVMs with word and character n-grams as features weighting each feature using TF-IDF. This indicates that simple methods are still competitive and far from being fully overcome by neural approaches.

PoliticIT has been fully run on CodaLab. The task is now in the Post-evaluation phase and it is still open to submissions from other teams, making this resource freely available to the NLP community for research purposes.

Future work will investigate the addition of an extra subtask related to stance detection, to determine which authors are in favor of certain topics and which users are against. We can use this information to define clusters of users and to observe whether there is a relationship between the topics and the political ideology.

Acknowledgments

This work is part of the research projects LaTe4PoliticES (PID2022-138099OB-I00) funded by MCIN/AEI/10.13039/501100011033 and the European Fund for Regional Development (FEDER)-a way to make Europe and LaTe4PSP (PID2019-107652RB-I00/AEI/ 10.13039/501100011033) funded by MCIN/AEI/10.13039/501100011033. This work is also part of the research projects AIInFunds (PDC2021-121112-I00) and LT-SWM (TED2021-131167B-I00) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. It also has been partially supported by Project CONSENSO (PID2021-122263OB-C21), Project MODERATES (TED2021-130145B-I00) and Project SocialTox (PDC2022-133146-C21) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, Project PRECOM (SUBV-00016) funded by the Ministry of Consumer Affairs of the Spanish Government, Project FedDAP (PID2020-116118GA-I00) supported by MICINN/AEI/10.13039/501100011033 and WeLee project (1380939, FEDER Andalucía 2014-2020) funded by the Andalusian Regional Government. Salud María Jiménez-Zafra has been partially supported by a grant from Fondo Social Europeo and the Administration of the Junta de Andalucía (DOC_01073).

References

[1] N. Fairclough, Critical discourse analysis: The critical study of language, Routledge, 2013.

[2] B. Verhulst, L. J. Eaves, P. K. Hatemi, Correlation not causation: The relationship between personality traits and political ideologies, American journal of political science 56 (2012) 34–51.

[3] M. Fatke, Personality traits and political ideology: A first global assessment, Political Psychology 38 (2017) 881–899.

[4] B. Baumgaertner, J. E. Carlisle, F. Justwan, The influence of political ideology and trust on willingness to vaccinate, PloS one 13 (2018) e0191728.

[5] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, EVALITA 2023: Overview of the 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, in: M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi (Eds.), Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[6] J. Bevendorff, B. Chulvi, G. L. D. L. Peña Sarracén, M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. Rangel, P. Rosso, et al., Overview of PAN 2021: authorship verification, profiling hate speech spreaders on twitter, and style change detection, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2021, pp. 419–431.

[7] J. A. García-Díaz, S. M. Jiménez-Zafra, M.-T. M. Valdivia, F. García-Sánchez, L. A. Ureña-López, R. Valencia-García, Overview of PoliticEs 2022: Spanish Author Profiling for Political Ideology, Procesamiento del Lenguaje Natural 69 (2022) 265–272.

[8] J. A. García-Díaz, Á. Almela, G. Alcaraz-Mármol, R. Valencia-García, Umucorpusclassifier: Compilation and evaluation of linguistic corpus for natural language processing tasks, Procesamiento del Lenguaje Natural 65 (2020) 139–142.

[9] J. A. García-Díaz, R. Colomo-Palacios, R. Valencia-García, Psychographic traits identification based on political ideology: An author analysis study on spanish politicians' tweets posted in 2020, Future Generation Computer Systems 130 (2022) 59–74.

[10] A. Joulin, É. Grave, P. Bojanowski, T. Mikolov, Bag of tricks for efficient text classification, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, 2017, pp. 427–431.

[11] M. Grootendorst, Bertopic: Neural topic modeling with a class-based tf-idf procedure, arXiv preprint arXiv:2203.05794 (2022).

[12] S. Schweter, Italian bert and electra models, 2020. URL: https://doi.org/10.5281/zenodo.4263142. doi:10.5281/zenodo.4263142.

[13] C. D. Hromei, D. Croce, V. Basile, R. Basili, ExtremITA at EVALITA 2023: Multi-Task Sustainable Scaling to Large Language Models at its Extreme, EVALITA 2023 Eigth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (2023) –.

[14] A. Santilli, Camoscio: An italian instruction-tuned llama, https://github.com/teelinsan/camoscio, 2023.

[15] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T. B. Hashimoto, Stanford alpaca: An instruction-following llama model, https://github.com/tatsu-lab/stanford_alpaca, 2023.

[16] G. Sarti, M. Nissim, It5: Large-scale text-to-text pretraining for italian language understanding and generation, ArXiv preprint 2203.03759 (2022). URL: https://arxiv.org/abs/2203.03759.

[17] H. Cabrera-Pineda, E. S. Téllez, S. Miranda, INFOTEC-LaBD at PoliticIT: Political Ideology Detection in Italian Texts, EVALITA 2023 Eigth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (2023) –.

[18] C. Çöltekin, M. Brivio, F. Can, Tübingen at PoliticIT: Exploring SVMs, Pretrained Language Models, and Linguistic Transfer for Ideology Detection in Social Media, EVALITA 2023 Eigth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (2023) –.

[19] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, É. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440–8451.

[20] T. Erjavec, M. Ogrodniczuk, P. Osenova, N. Ljubešić, K. Simov, A. Pančur, M. Rudolf, M. Kopp, S. Barkarson, S. Steingrímsson, et al., The parlamint corpora of parliamentary proceedings, Language resources and evaluation 57 (2023) 415–448.

[21] R. Pan, Á. Almela, F. García-Sánchez, UMUTeam at PoliticIT-EVALITA2023: Evaluating Transformer Model for Detecting Political Ideology in Italian Texts, EVALITA 2023 Eigth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (2023) –.

[22] M. Á. Rodríguez-García, URJC-Team at EVALITA 2023: Political Ideology Detection in Italian Texts Using Transformers Architectures, EVALITA 2023 Eigth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (2023) –.

[23] J. Tiedemann, Parallel data, tools and interfaces in OPUS, in: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey, 2012, pp. 2214–2218. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.

[24] P. J. Ortiz Suárez, B. Sagot, L. Romary, Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures, Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, Leibniz-Institut für Deutsche Sprache, Mannheim, 2019, pp. 9 – 16. URL: http://nbn-resolving.de/urn:nbn:de:bsz:mh39-90215. doi:10.14618/ids-pub-9021.

[25] M. Polignano, P. Basile, M. de Gemmis, G. Semeraro, V. Basile, AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets, in: Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019), volume 2481, CEUR, 2019. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85074851349&partnerID=40&md5=7abed946e06f76b3825ae5e294ffac14.

INFOTEC-LaBD at PoliticIT: Political Ideology Detection in Italian Texts

Hiram Cabrera-Pineda[1,*], Eric Sadit Téllez[1,2,3] and Sabino Miranda[1,3,4]

[1]INFOTEC Aguascalientes, México
[2]CICESE, Ensenada, Baja California, México
[3]CONAHCYT, México
[4]UPIITA-IPN, Ciudad de México, México

Abstract

This working notes presents our approach to clusters of texts profiling in the PoliticIT 2023 challenge, utilizing a non-linear low-dimensional representation of term distribution entropy. Our proposed algorithm is designed to learn a 3-dimensional model of text data, effectively capturing essential features for accurate profiling. Furthermore, it offers valuable insights through cluster analysis and visualizations, enabling a deeper understanding of the underlying patterns. The method employed in our algorithm uses a bag-of-words representation and incorporates weighting schemes based on the term's distribution entropy. By leveraging these techniques, we are able to extract meaningful information and uncover significant characteristics related to clusters of texts profiling. To evaluate the effectiveness of our proposed algorithm, we conducted experiments on the PoliticIT 2023 dataset, encompassing three tasks: gender identification, and binary and multiclass political ideology classification. The obtained results demonstrate the competitiveness of our solution across all three tasks, highlighting its efficacy in accurately predicting the attributes of clusters of texts. One notable advantage of our algorithm is its explainability. It offers insights into the reasoning behind its predictions, allowing users to understand the factors influencing cluster of text behavior. This transparency enhances the practicality and utility of our algorithm as a powerful tool for cluster of text profiling and behavior analysis.

Keywords

explainable user-profiling, low-dimensional representations, political parties identification, term's distribution entropy weighting

1. Introduction

User profiling is a powerful technique used to extract valuable information about individuals, including their interests, demographics, behavioral patterns, and even their political preferences. Cluster profiling aims to identify patterns, trends, and similarities within groups of texts written by different users with similar traits or characteristics based on specific criteria. By carefully analyzing data related to users, we can gain deep insights into their unique characteristics and preferences. These insights have far-reaching applications in various domains, including enhancing user experiences, personalized advertising, detecting and preventing malicious activities, and gaining a deeper understanding of societal dynamics and preferences. In this manuscript, our focus lies specifically on political preferences as participation in the PoliticIT challenge of the EVALITA 2023 forum.

The task consists of predicting user demographics like gender and political ideology from 36,240 Twitter messages in the Italian language.These tweets cover various topics, from news and current events to personal thoughts and experiences. Twitter is a valuable data source for user profiling, offering valuable insights into users' interests, demographics, and behaviors. The practical applications of these tasks span various domains, including personalized marketing, content recommendation, and social science research [1].

Automatic user profiling gives us insight into a population's characteristics, preferences, and ideological orientations. There are several examples of this, including the PAN@CLEF and FIRE series. These tools cover a wide range of objectives such as age, gender, language variety identification, and personality recognition in different languages and genres, like blogs, reviews, social media, and Twitter [2, 3, 4, 5, 6]. Forums like MEX-A3T [7] delve into identifying occupation types and places of residence. At the same time, the IberLEF@SEPLN forum focuses on gender, profession, and political ideology profiling, see [8, 9], all of which contribute significantly to the advancement of user profiling techniques.

This paper provides an overview of the author profiling task, focusing on its relevance in political domains, specifically in the context of the PoliticIT challenge. Our approach is described in Section 2, where we outline the methodology we employed. In Section 3, we delve into

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
hiram.cabrerap@infotec.mx (H. Cabrera-Pineda); eric.tellez@ieee.org (E. S. Téllez); smiranda@ieee.org (S. Miranda)
0000-0003-1419-761X (H. Cabrera-Pineda); 0000-0001-5804-9868 (E. S. Téllez); 0000-0002-9421-8566 (S. Miranda)
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper8.pdf

the details of our dataset and models, providing a brief analysis. Section 4 presents the experimental results and their discussion. Finally, in Section 5, we reflect upon the achievements of our approach and discuss their implications.

2. Our approach

Our primary objective is to develop robust and interpretable representations that achieve high performance and allow for in-depth analysis of the model's label predictions. Instead of relying on a predefined set of features, we generate 3D maps that capture the underlying similarity structure of the original high-dimensional representation. This spatial representation lets us explore similar clusters of texts by directly examining their messages and analyzing their shared vocabulary.

In a previous publication [10], we introduced our overarching approach. To tackle the PoliticIT 2023 challenge, we have streamlined our approach by reducing the number of hyperparameters. Our cluster of texts profiling methodology encompasses three main modules, providing a more concise and focused framework.

Vector spaces models. We implemented various preprocessing steps to prepare the data; for instance, we converted all messages to lowercase, normalized blank spaces to a single space, and removed diacritic marks. Token numbers 1-9 were preserved to capture important information on small numbers, while other numbers were replaced by 0 (to reduce dimensionality). We employed three types of tokens: unigrams, bigrams, and character q-grams of size four. Each token is modeled as a distribution along classes, and we compute the token's weight based on the distribution's entropy using the formulation of [10], that is, for each token t

$$\mathsf{entweight}(t) = 1 + \frac{\sum_{c \in L} p_c^t \log p_c^t}{\log \# L}$$

where L is the set of tables and p_c^t is the probability of token t in class c. Note that the numerator's $\log$ produces negative numbers. Therefore, the weight is bounded between 0 (tokens with low-discrimination power) and 1 (high-discriminant tokens). This formulation ignores smoothing constants as needed by our original formulation. Instead, we reject tokens from the vocabulary if they do not occur in at least M clusters of texts. This change reduces the memory required by our models and speedups computations.

Non-linear dimensional reduction. The non-linear dimensional reduction module uses the Uniform Manifold Approximation and Projection (UMAP) [11] to produce a low-dimensional vector space (UMAP model) from the resulting vector space in the previous Module. We made three-dimensional projections of the data for visualization and as the input of classifiers. UMAP projection requires the k nearest neighbor graph (using our high-dimension vector space and cosine similarity) to capture the dataset's structure. More detailed, a fuzzy smoothed version of the graph is created, and then, the eigenvectors of the normalized Laplacian matrix are computed to initialize low-dimensional embedding vectors that are finally optimized to preserve the k nearest neighbor graph in the low-dimensional projection. This procedure captures similar properties to the spectral clustering [12] while producing insightful visualization. The explainability aspect utilizes UMAP projections to create visualizations for understanding data distribution, cluster relationships, and separability. These visualizations provide intuitive representations of clusters, highlighting proximity, separability, and discernible patterns or trends.

Classification. The supervised learning stage uses the low-dimensional vectors to train a classifier to predict the cluster's political ideology, or gender. We use SVM classifiers with linear and non-linear kernels in the sklearn library [13]. We perform a model selection procedure for tuning each classifier.

3. Methodology and model analysis

This manuscript addresses the challenge of profiling political preferences within the context of the PoliticIT challenge held in the EVALITA 2023 forum. The task involves predicting user gender and political ideology based on a dataset comprising more than 36 thousand Twitter messages written in Italian. To ensure privacy and ethical considerations, an automated clustering approach was employed, grouping tweet messages from different users who shared all the evaluated traits [14].

The provided training corpus consists of 103,840 messages collected from Twitter, aiming to extract the author's traits from Italian texts. This shared task entails gathering demographic traits, i.e., gender and political ideology as psychographic traits. See [14] for more details.

The training dataset includes tweets from 1,298 clusters of texts, each contributing 80 tweets. However, it is important to note that the dataset exhibits varying degrees of class imbalance across different class categories. Considering the gender task, the proportion is 63.4% vs. 37.6% for males and females, respectively. There are 12.5%, 43%, 10.1%, and 34.4% between left, moderate left, moderate right, and right regarding multiclass political ideology. The proportion is 55.4% vs. 44.5% for

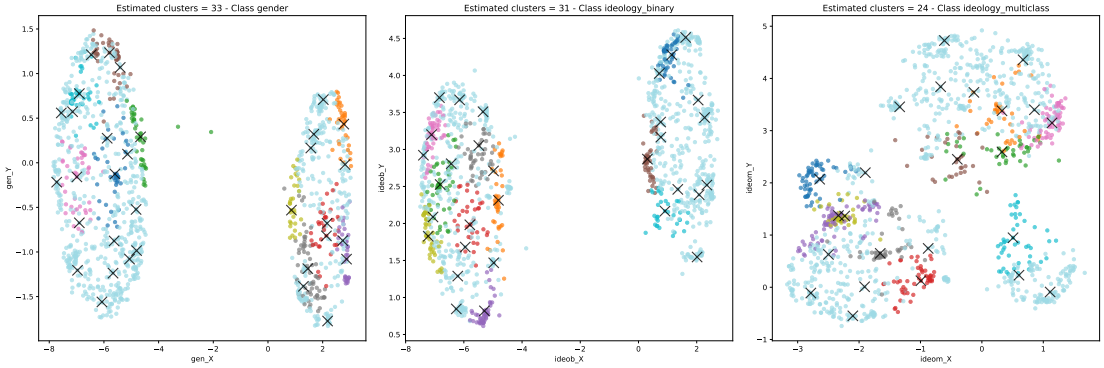


Figure 1: 2D UMAP projections for each task with $M = 5$ and $k = 50$.

binary ideology, i.e., a slight imbalance between the left and right categories, respectively. We are asked to create models to predict these labels for a test dataset, i.e., a list of 453 clusters (80 messages per cluster); the label distribution of the test dataset is unknown.

As explained in Section 2, we create low-dimensional projections that concisely represent the dataset and its associated labels. This visualization serves the purpose of uncovering group properties and revealing any underlying cluster structures. It is important to note that our method utilizes a k nearest neighbor graph constructed from a vector space generated by the entropy weighting model. To ensure meaningful results, we exclude tokens with low frequency, specifically those that appear in fewer than M clusters. Hence, our primary hyperparameters revolve around k and M.

We experimented with different parameter values to create UMAP low-dimensional projections. Specifically, we varied the values of k and M, with k ranging from 10 to 50 and M ranging from 3 to 35. We generate three-dimensional embeddings using a spectral layout for embedding initialization and optimized during 100 epochs. We also used three negative samples per point (user vector).

Based on the combinations of k and M parameters, we obtained 35 unique UMAP projections for each task in the challenge, including gender, binary ideology, and multiclass ideology. These projections provided visual representations that captured the data's underlying structure and relationships.

In Figure 1, we present an overview of different parameter combinations (k and M) and their impact on the data representation. Please note that although our models use three-dimensional projections, the figures shown are two-dimensional for easier visualization. The figures display centers obtained through the Affinity Propagation (AP) clustering algorithm [15] applied to the low-dimensional embeddings. These results offer insights into patterns, separations, and overlaps within the data, aiding the analysis and interpretation. It's important to clarify that the centers represent groups of near clusters of texts, not individual classes.

We used the homogeneity score to evaluate the extent to which the clusters are homogeneous, meaning that the samples within each group are similar to each other [16]. It measures the similarity of the clusters in terms of their composition and whether they predominantly consist of samples from a single ground truth class. A high homogeneity score indicates that the clusters formed by the AP algorithm are internally consistent. However, it's important to note that the homogeneity score does not directly measure how well the predicted clusters match the ground truth labels. It doesn't assess whether the cluster labels perfectly align with the true class labels. Instead, it assesses the consistency and purity of the groups within themselves.

We provided homogeneity heatmaps in Figure 2. They use color intensity to represent the homogeneity scores for different combinations of k and M. Darker colors indicate higher homogeneity.

By examining the homogeneity score heatmaps, we can gain insights into the clustering quality and make better decisions regarding the choice of parameters k and M. This helps us understand how different parameters impact the quality of the clusters.

3.1. Model Selection

Our participation in the PoliticIT challenge involved the development of multiple models using the provided training data. To select the most suitable models, we employed a Grid Search approach and conducted model selection based on maximizing the macro-F1 score. The evaluation process utilized five-fold stratified cross-validation.

For the creation of our classification models, we utilized SVM classifiers with both linear and non-linear ker-

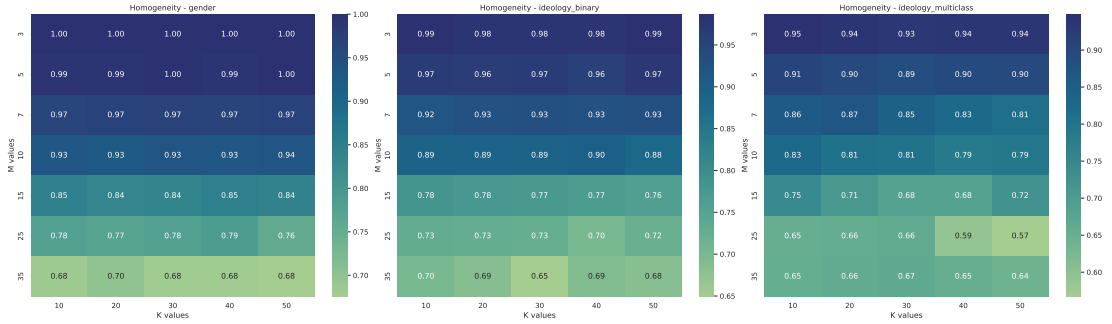


Figure 2: Homogeneity heatmaps for evaluation of low-dimensional clustering.

nels. Through a hyperparameter optimization process and 5-fold cross-validation, we identified the models that exhibited the best performance.

Each model corresponds to a distinct cluster of texts representation defined by parameters k and M. Additionally, we considered three-dimensional representations and the concatenation of the three 3D maps from all tasks, resulting in a 9-dimensional vector space.

Through this comprehensive evaluation, we thoroughly assessed the performance of various models and determined the most effective approaches for each task. The final selection of models was based on the parameter combination that achieved the highest accuracy during cross-validation.

4. Experimental results

In this section, we present the experimental results of our approach for the EVALITA 2023 PoliticIT task. The source code can be found at the following GitHub repository: https://github.com/hiramcp/PoliticIT2023.

Our experiments were conducted on a Windows 10 operating system, utilizing a four-core Laptop with 32 GB of RAM and an Intel Core i7-1165G7 @ 2.80GHz processor. To compute the vector space and perform preprocessing functions, tokenization, and entropy-based weighting, we utilized *TextSearch.jl* Julia package available at https://github.com/sadit/TextSearch.jl. For UMAP projections, we employed the *SimSearchManifoldLearning.jl* Julia package found at https://github.com/sadit/SimSearchManifoldLearning.jl. Lastly, model selection and classification tasks were carried out using the Python scikit-learn package [13].

We explored different values of k and M to find the right balance between preserving local and global structures. Analyzing homogeneity scores and heatmaps, we identified parameter combinations of k and M that yielded reliable and competitive projections for each task. We aimed to ensure that the data representations used for training and classification showed clear and meaningful clusters, increasing the likelihood of achieving competitive results in the training phase.

Table 1 shows each task's best values of k and M. We determined these combinations by evaluating projections using homogeneity scores and affinity propagation.

Task	k	M
Gender	30	7
Ideol. Binary	50	3
Ideol. Multiclass	50	5

Table 1: Optimal combinations of k ($k$nn graph) and M (minimum occurrence) parameters for each task in the challenge.

4.1. Prediction results

For the final selection, we chose the parameter combination for each task that achieved the highest macro F1 score during cross-validation, as described in §3. The best hyperparameters used for each task are summarized in Table 2.

Task	Classifier	Dim.	Hyperparameters
Gender	Linear SVM	3-D	no standardized, C=1000, class_weight=None, dual=False, max_iter=12000, penalty=L2, random_state=42
Ideol. Binary	Linear SVM	9-D	no standardized, C=1, class_weight=balanced, dual=False, max_iter=12000, penalty=L2, random_state=42
Ideol. Multiclass	RBF SVM	9-D	no standardized, C=100, class_weight=balanced, gamma=scale, kernel=rbf

Table 2: Best hyperparameters for the different user profiling model tasks.

We used the top-performing Machine Learning models from our previous tests to analyze the demographic and psychographic characteristics of Tweets in the Test dataset. We specifically aimed to identify self-reported gender as demographic trait, as well as political ideology as a psychographic trait.

Our approach attained an overall score of 0.800788 in the final evaluation, placing us in the 2nd position on the leaderboard. The individual task results are presented in Table 3.

Task	**Macro F1 Score (rank)**
Gender	0.824287 (1)
Ideology Binary	0.860230 (5)
Ideology Multiclass	0.717849 (2)

Table 3: Performance for the most suitable models on the final test dataset.

The macro-average score represents the overall accuracy of our models in predicting the assigned gender and political ideology using the Test dataset.

5. Conclusions

Our framework combines our entropy-based weighting scheme and non-linear dimensional reduction techniques to achieve a practical tradeoff between the model's introspection and high-quality predictions. By experimenting with various combinations of parameters for k and M, we gained valuable insights into task clustering and separation. The results emphasized the significance of considering the neighborhood size and the minimum number of documents required for identifying and distinguishing gender and ideology groups.

Future studies can explore different parameter configurations to assess their effects and a comprehensive error analysis can be conducted to gain valuable insights into the limitations of the models and identify areas for improvement. By identifying specific challenges or patterns in misclassifications, refined models can be developed, and targeted strategies can be implemented to address these issues.

It's also worth exploring other advanced dimensionality reduction techniques to enhance classification performance. Finally, our current approach is limited to lexical features; other representations (e.g., transformer-based ones) can help improve our approach in distinct scenarios.

Acknowledgments

H. Cabrera-Pineda is grateful to CONAHCyT for "Becas Nacionales para Estudios de Posgrado" scholarship No. 899981.

References

[1] C. I. Eke, A. A. Norman, L. Shuib, H. F. Nweke, A survey of user profiling: State-of-the-art, challenges, and solutions, IEEE Access 7 (2019) 144907–144924. doi:10.1109/ACCESS.2019.2944243.

[2] E. Stamatatos, M. Potthast, F. Rangel, P. Rosso, B. Stein, Overview of the pan/clef 2015 evaluation lab, in: Proceedings of the 6th International Conference on Experimental IR Meets Multilinguality, Multimodality, and Interaction - Volume 9283, CLEF'15, Springer-Verlag, Berlin, Heidelberg, 2015, p. 518–538. URL: https://doi.org/10.1007/978-3-319-24027-5_49. doi:10.1007/978-3-319-24027-5_49.

[3] F. M. R. Pardo, P. Rosso, M. M. y Gómez, M. Potthast, B. Stein, Overview of the 6th author profiling task at pan 2018: Multimodal gender identification in twitter, in: CLEF, 2018.

[4] P. Rosso, F. Rangel, Author profiling tracks at fire, SN Computer Science 1 (2020) 72. URL: https://doi.org/10.1007/s42979-020-0073-1. doi:10.1007/s42979-020-0073-1.

[5] J. Bevendorff, B. Chulvi, G. L. D. L. P. Sarracén, M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, , E. Zangerle, Overview of PAN 2021: Authorship Verification,Profiling Hate Speech Spreaders on Twitter,and Style Change Detection, in: 12th International Conference of the CLEF Association (CLEF 2021), Springer, 2021.

[6] F. Rangel, G. L. D. L. P. Sarracén, B. Chulvi, E. Fersini, P. Rosso, Profiling Hate Speech Spreaders on Twitter Task at PAN 2021, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.

[7] M. Á. Álvarez-Carmona, E. Guzmán-Falcón, M. Montes-y Gómez, H. J. Escalante, L. Villaseñor-Pineda, V. Reyes-Meza, A. Rico-Sulayes, Overview of mex-a3t at ibereval 2018: Authorship and aggressiveness analysis in mexican spanish tweets, in: Notebook Papers of 3rd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL), Seville, Spain, September, 2018.

[8] J. A. García-Díaz, S. M. Jiménez-Zafra, M. T. Martín-Valdivia, F. García-Sánchez, L. A. Ureña-López, R. Valencia-García, Overview of PoliticES 2022: Spanish Author Profiling for Political Ideology, Procesamiento del Lenguaje Natural 69 (2022).

[9] J. A. García-Díaz, S. M. Jiménez-Zafra, M. T. Martín-Valdivia, F. García-Sánchez, L. A. Ureña-López, R. Valencia-García, Overview of PoliticES 2023 at IberLEF: Political Ideology Detection in Spanish Texts, Procesamiento del Lenguaje Natural 71 (2023).

[10] H. Cabrera, E. S. Téllez, S. Miranda, Infotec-labd at politices 2022: Low-dimensional stacking model for political ideology profiling, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022). CEUR Workshop Proceedings, CEUR-WS, A Coruna, Spain, 2022.

[11] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, 2018. URL: https://arxiv.org/abs/1802.03426. doi:10.48550/ARXIV.1802.03426.

[12] A. Ng, M. Jordan, Y. Weiss, On spectral clustering: Analysis and an algorithm, Advances in neural information processing systems 14 (2001).

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[14] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[15] B. J. Frey, D. Dueck, Clustering by passing messages between data points, science 315 (2007) 972–976.

[16] A. Rosenberg, J. Hirschberg, V-measure: A conditional entropy-based external cluster evaluation measure, in: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning(EMNLP-CoNLL), 2007, pp. 410–420.

UMUTeam at PoliticIT-EVALITA2023: Evaluating Transformer Model for Detecting Political Ideology in Italian Texts

Ronghao Pan[1], Ángela Almela[2] and Francisco García-Sánchez[1]

[1]Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100, Spain
[2]Facultad de Letras, Universidad de Murcia, Campus de La Merced, 30001, Murcia, Spain

Abstract

This paper describes the participation of the UMUTeam in the PoliticIT shared task organized at EVALITA 2023. It is an automatic document classification task on clusters of texts, which consists of extracting self-assigned gender as a demographic trait, and ideology as a psychographic trait through a set of texts written in Italian by several authors sharing these traits. For this task, we used the fine-tuning approach of a pre-trained transformer-based masked language model for Italian called *dbmdz/bert-base-italian-cased* to carry out the identification of different features. After several submissions for these tasks, our team ranked sixth out of 7 participants, with an average F1 score of 70.426% of all classification models. However, our binary political ideology classification model obtained the fourth-best result with an F1 score of 86.63%.

Keywords

Natural Language Processing, Transformers, Politic ideology detection, Large Language Model, Multiclass classification

1. Introduction

Political ideology is considered to be a psychographic trait that can be used to understand individual and social behavior, including moral and ethical values, as well as inherent attitudes, appraisals, biases, and prejudices [1]. The relationship between personality traits and political ideology was demonstrated in [2], in which the authors collected data from 21 countries and concluded that political ideology is related to the big five personality traits and that the results of this relationship vary across countries, especially as a function of the level of prosperity.

Political ideology has a great influence on society and can guide our daily decisions, just like other psychographic traits such as personality. However, these decisions are made both consciously and unconsciously, as the social and cultural environment influences our ideology. In general, however, people tend to be more reluctant to follow advice and directions from politicians who do not coincide with their ideology [3]. In some extreme cases, people may be strongly biased in favor of one political party and, at the same time, strongly disagree with others, which may lead to irrational decisions and put people's lives at risk by ignoring specific recommendations of the authorities.

Thus, the aim of PoliticIT at EVALITA 2023 [4] is to extract information about users' political ideology through text clusters in Italian. For this purpose, the task is divided into several objectives such as extracting self-assigned gender as a demographic trait, and ideology as a psychographic trait through a set of texts written in Italian by several authors sharing these traits. This task builds on a previous task called PoliticES presented at IberLEF 2022 [5] where the dataset was an extension of the PoliCorpus 2020 dataset [3].

Transformer models are a neural network architecture used in various fields of NLP and other machine learning domains. The idea of a Transformer network originated from a paper called *Attention is all you need* [6], developed by researchers at Google to address the language translation problem. Unlike recurrent neural networks (RNNs) and convolutional neural networks (CNNs), which were widely used in NLP before the advent of Transformers, this model relies on attention mechanisms to process sequences more efficiently and capture long-range relationships between words in a sentence or the order of words in a sequence [7]. Pretrained models based on Transformers are language models generated through massive training on large amounts of unlabeled text, aiming to learn language patterns and features from the training data. This enables them to capture semantic, syntactic, and contextual information, developing a deep understanding of language in general. For example, BERT (Bidirectional Encoder Representations from Transformers) is a Transformer-based model that achieved remarkable accuracy and advanced the state of the art in various Natural Language Processing (NLP) tasks [8]. Another advantage of pre-trained models is the ability to add spe-

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
ronghao.pan@um.es (R. Pan); angelalm@um.es (Á. Almela); frgarcia@um.es (F. García-Sánchez)
0009-0008-7317-7145 (R. Pan); 0000-0002-1327-8410 (Á. Almela); 0000-0003-2667-5359 (F. García-Sánchez)
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper9.pdf

cific layers to adapt them to particular tasks. In other words, the process of taking a pre-trained model on vast amounts of data and continuing the training on a smaller and task-specific dataset is known as fine-tuning. The advantage of fine-tuning is that it allows leveraging the general language knowledge acquired during pretraining and applying it to specific tasks with relatively small amounts of training data.

This paper presents the UMUTeam's participation in this shared task, based on the fine-tuning of a pre-trained Italian Transformer model to detect the gender and political ideology of users through their texts written in Italian (document-level classification). The rest of the paper is organized as follows. Section 2 presents the task and dataset provided. Section 3 describes the methodology of our proposed system for addressing the task. Section 4 shows the results obtained. Finally, Section 5 concludes the paper with some findings and possible future work.

2. Task description

The shared task *PoliticIT* 2023, organized at EVALITA [9], is a document classification problem that aims to extract self-assigned gender as a demographic trait, and political ideology as a psychographic trait from a given text cluster [4]. The organizers propose political ideology as a binary classification problem and a multi-class classification problem. The organizers created text clusters by mixing some of these extracted tweets to avoid ethical and privacy concerns about author profiling on Twitter. Consequently, all clusters are composed of texts written by different users sharing all the features under evaluation. The users of the dataset are labeled with their gender (male, female), and political spectrum on two axes: binary (left, right) and multi-class (left, moderate left, moderate right, right). Regarding the tweets collected from each user, the organizers discarded retweets and tweets that contain headlines from news sites and removed tweets written in languages other than Italian. Moreover, they anonymized them by replacing all mentions of the politicians and other Twitter accounts mentions with *@user*. Furthermore, other entities, such as political party references, are also replaced with *@political_party* token. The tweets that belong to each cluster are selected favoring diversity, including texts from different dates and topics. For this shared task, the dataset is divided into training and test sets (80%-20%). For these different classification problems, we have used the same approach, which consists in fine-tuning a pre-trained Italian model based on Transformer for different demographic and psychographic trait classification tasks at the tweet level. Thus, the training set consists of 103,840 tweets from 1,298 clusters and the test set consists of 453 users with a total of 36,240 tweets. The distribution of demographic and psychographic features of the cluster is shown in Table 1.

Trait		Training	Test	Total
Gender	**Male**	810	318	**1128**
	Female	488	135	**623**
Ideology binary	**Left**	720	248	**968**
	Right	578	205	**783**
Ideology multi-class	**Moderate left**	558	148	**706**
	Left	162	100	**262**
	Moderate right	131	154	**285**
	Right	447	51	**498**

Table 1
The distribution of demographic and psychographic traits of the clusters.

3. Methodology

This task involves identifying the gender (demographic trait) and political ideology (psychographic trait) of users in a given set of texts (document-level classification). The task addresses gender traits as a binary classification problem and political ideology as both a binary and a multi-class classification problem. The approach used to solve the challenge of classifying both demographic and psychographic features at the document level consists in creating a phrase-level classification model by fine-tuning a pre-trained Italian model based on BERT [8] called *dbmdz/bert-base-italian-cased* [10]. The system architecture is depicted in Figure 1 and the pipeline used to participate in this task can be described as follows. First, the dataset has been processed, and the emoji in the text have been transformed into text through the *emoji* library[1]. Second, the training set has been divided into two parts: training and validation. Third, classification models are created for each feature using the approach explained above. Fourth, having the classification models at the sentence level, two strategies have been evaluated to identify the demographic and psychographic traits of the users (at document level): (1) **mode**, which consists in predicting each user's tweet individually and selecting the most repeated label among the results obtained with the classifier, and (2) **highest probability**, which selects the label with the highest probability. Next, some more details about the preprocessing and fine-tuning stages are provided.

3.1. Dataset preprocessing

Tweets from the user have the same demographic and psychographic features, so to carry out the task of identifying these features, we created a sentence-level classification model for each feature with all tweets from all users as the training set. The distribution of demographic

[1] https://github.com/kyokomi/emoji

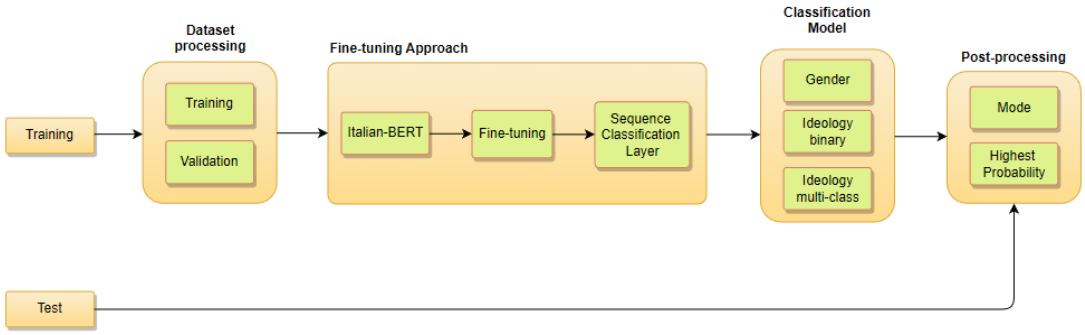


Figure 1: Overall system architecture.

and psychographic features in the dataset is shown in Table 2. In this case, the organizers have only provided the training set, so we have divided the training set into two subsets (80-20%): training and validation. The custom validation split is created using stratified sampling, in order to maintain a balance between labels.

Trait		Training	Validation
Gender	**Male**	51840	12960
	Female	31232	7808
Ideology binary	**Left**	46080	11520
	Right	36992	9248
Ideology multi-class	**Moderate left**	35712	8928
	Left	10368	2592
	Moderate right	8384	2096
	Right	28608	7152

Table 2
The distribution of demographic and psychographic traits of the training and validation split.

3.2. Fine-tuning approach

We utilized the fine-tuning approach of a transformer-based masked language model for Italian called *dbmdz/bert-base-italian-cased*[2] to carry out the identification of different features. Based on the BERT base model, *dbmdz/bert-base-italian-cased* has been pre-trained using a recent Wikipedia dump and various texts from the OPUS corpora collection [11]. The final training corpus has a size of 13GB and 2,050,057,573 tokens. The fine-tuning process consists of adapting and adding a classification layer to the model to perform the training of the full model. In this way, the model takes advantage of the pre-trained linguistic knowledge of *dbmdz/bert-base-italian-cased* and adapts it specifically for a particular classification task, which can significantly improve performance on that task. To do this, we have used the "transformers" library[3] from HuggingFace, which includes a classification layer for pre-trained models. This classification layer is responsible for mapping the contextualized representations of input tokens to the specific output of the corresponding task. For sequence classification, the output layer usually consists of a classification head, which comprises one or more fully connected layers followed by a softmax activation. Finally, the model has been fine-tuned with a training batch size of 16, 6 epochs, a learning rate of 2e-5, and a weight decay of 0.01.

4. Results

This section describes the systems presented by our team in each run and the overall results obtained in this shared task. It should be noted that each participating team could submit ten runs.

We submitted a total of 2 runs for this task. The results and a brief description of each are shown in Table 3. The first run is based on running the different classification models on the user texts and then using the "mode" strategy, which consists in selecting the most frequent label obtained in the set of user texts for each feature. This run obtained an average F1 score of 0.70426. The second run has the same structure as the first one but with a different post-processing strategy; in this case, the "high probability" strategy has been used, which consists of making a decision based on the highest probability of all the predictions obtained on the user texts. In this case, it performed worse than with the "mode" strategy, with an average F1 score of 0.49232.

To perform the error analysis and check in which case the models give wrong predictions, a normalized confusion matrix with truth labels has been used, which consists of a table showing the distribution of the predictions of a model with respect to the truth label of the

[2]https://github.com/dbmdz/berts

[3]https://huggingface.co/docs/transformers/index

Strategy	Avg. F1	Gender			Ideology binary			Ideology multi-class		
		M-P	M-R	M-F1	M-P	M-R	M-F1	M-P	M-R	M-F1
Mode	**0.70426**	0.78736	0.69179	**0.71271**	0.87796	0.86244	**0.86638**	0.85372	0.55402	**0.53369**
Highest Prob.	**0.49323**	0.63904	0.62732	**0.63168**	0.68540	0.58355	**0.53658**	0.39979	0.36964	**0.31144**

Table 3
Results for each feature. For each strategy, the macro precision (M-P), macro recall (M-R), and macro F1-score (M-F1) are reported.

#	Team Name	Avg. F1	F1 Gender	F1 Ideology Binary	F1 Ideology Multi-class
1	Tübingen	0.824057	0.792469	0.928223	0.751477
2	INFOTEC-LaBD	0.800788	0.824287	0.860230	0.717849
3	extremITA	0.771867	0.769002	0.925572	0.621027
4	INGEOTEC	0.762001	0.732259	0.848525	0.705220
5	Teeeech	0.751067	0.752146	0.887720	0.613335
6	**UMUTeam**	**0.704262**	**0.712714**	**0.866382**	**0.533688**
7	NLP_URJC	0.684715	0.661327	0.770840	0.621979

Table 4
Official leaderboard PoliticIT shared task.

data. The confusion matrix of the system using the mode strategy is shown in Figure 2. It can be observed that our model tends to confuse the female gender with the male gender and right-wing with left-wing political ideology, and this may be due to the fact that there are fewer examples in the training set (see Table 2), which makes the model predict the wrong case. In the multi-class classification model of political ideology in our model, it fails mainly in the left-wing prediction, which tends to predict moderate left with a probability of 58%, and the moderate right-wing prediction, which tends to predict moderate left with a probability of 74.51%.

Figure 3 shows the confusion matrices of the system with the highest probability strategy. It can be observed that this strategy tends to confuse right and left ideology in the identification of binary ideology, so in Figure 3b it is shown that there are 78.05% cases that have been mispredicted. As for the identification of multiclass ideology, the model performs better than the mode strategy in identifying left and moderate right with a probability of 40% and 52.94%, respectively. However, it tends to misidentify moderate left and right ideology poorly, with hit probabilities of 9.45% and 45.45% and a difference of 89.86% and 40.91% with respect to the mode strategy (see Figure 2c and 3c).

The official leaderboard for this task is depicted in Table 4. We achieved the sixth position in the ranking with an average F1 score of 0.7042. However, our binary political ideology classification model obtained the fourth-best result with an F1 score of 0.8663.

5. Conclusion

This paper summarizes the participation of UMUTeam in the PoliticIT shared task (EVALITA 2023). We achieved a 6/7 on the mean of all F1 scores (70.42%) for the demographic and psychographic feature identification models. For this, we used the fine-tuning approach of a pre-trained transformer-based masked language model for Italian called *dbmdz/bert-base-italian-cased*[4] to carry out the identification of different features.

As future work, we are planning to improve our pipeline using an expanded transformer-based masked language model with political speech, i.e., extend a Masked Language Model (MLM) model with political text and later fine-tuning this model for detecting political ideology. In addition, we are planning to fine-tune other pre-trained Italian models to see if they improve the *dbmdz/bert-base-italian-cased* performance.

Acknowledgments

This work is part of the research projects LaTe4PoliticES (PID2022-138099OB-I00) funded by MCIN/AEI/10.13039/501100011033 and the European Fund for Regional Development (FEDER)-a way to make Europe and LaTe4PSP (PID2019-107652RB-I00/AEI/ 10.13039/501100011033) funded by MCIN/AEI/10.13039/501100011033. This work is also part of the research projects AIInFunds (PDC2021-

[4]https://github.com/dbmdz/berts

121112-I00) and LT-SWM (TED2021-131167B-I00) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR.

References

[1] B. Verhulst, L. J. Eaves, P. K. Hatemi, Correlation not causation: The relationship between personality traits and political ideologies, American Journal of Political Science 56 (2012) 34–51. doi:https://doi.org/10.1111/j.1540-5907.2011.00568.x.

[2] M. Fatke, Personality traits and political ideology: A first global assessment, Political Psychology 38 (2017) 881–899. doi:https://doi.org/10.1111/pops.12347.

[3] J. A. García-Díaz, R. Colomo-Palacios, R. Valencia-García, Psychographic traits identification based on political ideology: An author analysis study on spanish politicians' tweets posted in 2020, Future Generation Computer Systems 130 (2022) 59–74. doi:https://doi.org/10.1016/j.future.2021.12.011.

[4] Russo, Daniel and Jiménez Zafra, Salud María and García Díaz, José Antonio and Caselli, Tommaso and Guerini, Marco and Ureña López, Luis Alfonso, and Valencia García, Rafael, Overview of PoliticIT2023@EVALITA: Political Ideology Detection in Italian Texts, 8th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2023) (2023).

[5] García-Díaz, José Antonio AND Jiménez Zafra, Salud M. AND Martín Valdivia, María Teresa AND García-Sánchez, Francisco AND Ureña López, Luis Alfonso AND Valencia García, Rafael, Overview of politices 2022: Spanish author profiling for political ideology, Procesamiento del Lenguaje Natural 69 (2022a) 265–272.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, CoRR abs/1706.03762 (2017). URL: http://arxiv.org/abs/1706.03762. arXiv:1706.03762.

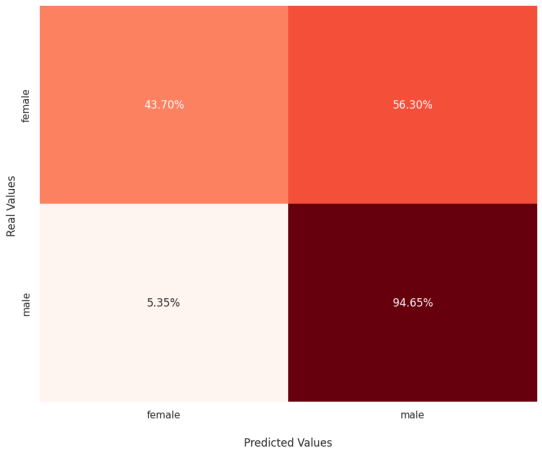
[7] T. Lin, Y. Wang, X. Liu, X. Qiu, A survey of transformers, AI Open 3 (2022) 111–132. URL: https://www.sciencedirect.com/science/article/pii/S2666651022000146. doi:https://doi.org/10.1016/j.aiopen.2022.10.001.

[8] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

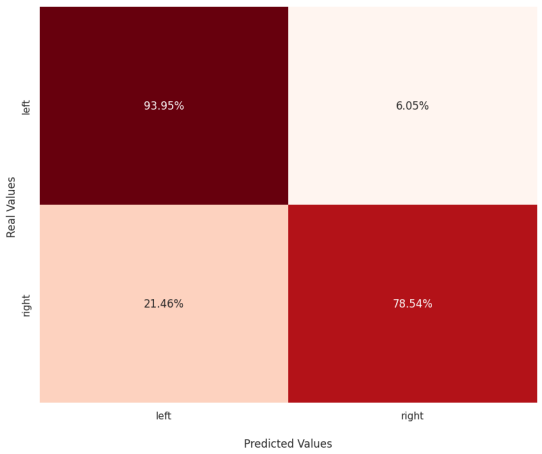
[9] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[10] S. Schweter, Italian bert and electra models, 2020. URL: https://doi.org/10.5281/zenodo.4263142. doi:10.5281/zenodo.4263142.

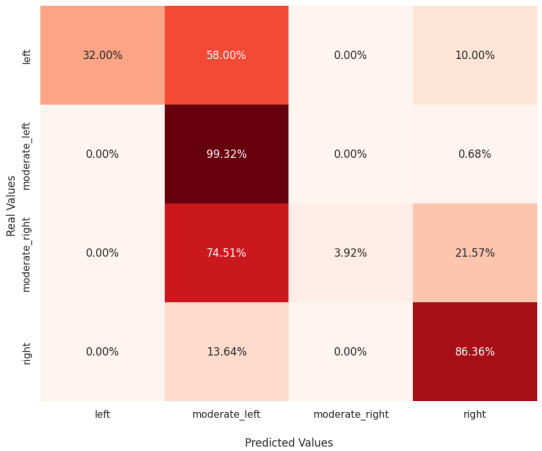
[11] J. Tiedemann, Parallel data, tools and interfaces in OPUS, in: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey, 2012, pp. 2214–2218. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.



(a) Gender

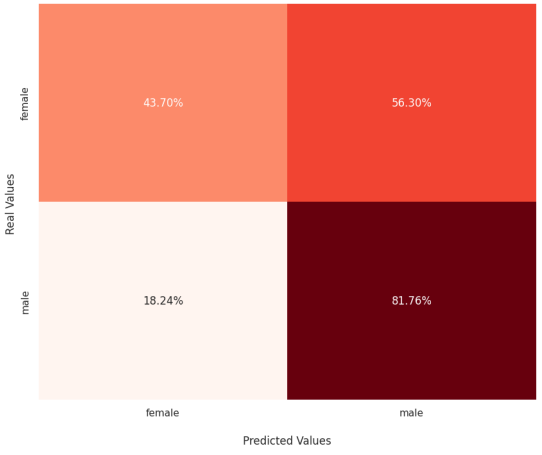


(b) Ideology binary

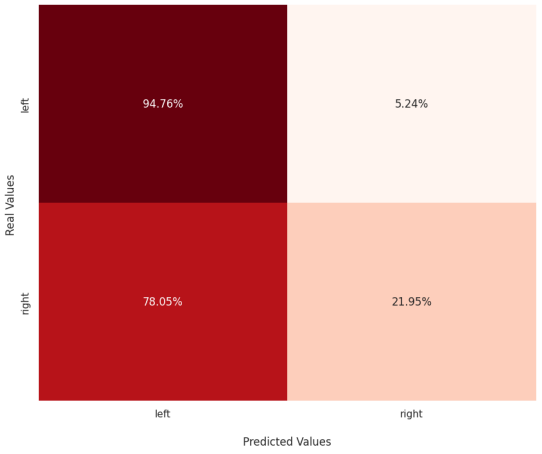


(c) Ideology multi-class

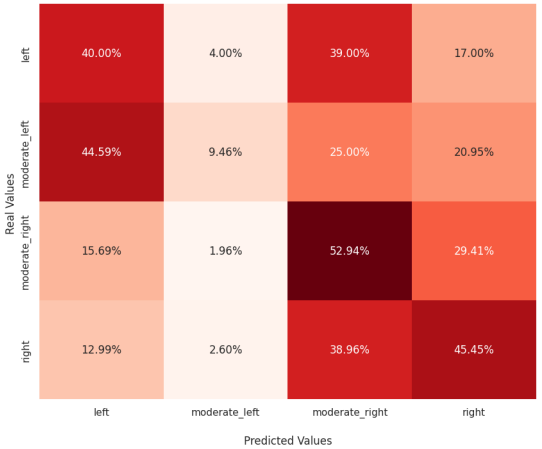
Figure 2: Confusion matrix of the system with mode strategy.



(a) Gender



(b) Ideology binary



(c) Ideology multi-class

Figure 3: Confusion matrix of the system with the highest probability strategy.

Tübingen at PoliticIT: Exploring SVMs, Pretrained Language Models, and Linguistic Transfer for Ideology Detection in Social Media

Çağrı Çöltekin[1], Matteo Brivio[1] and Fidan Can[1]

[1]University of Tübingen, Department of Linguistics

Abstract

This paper describes our approach to the EVALITA 2023 PoliticIT task on predicting political ideology and gender from Italian tweets. Furthermore, we investigate the effects of out-of-domain data (transcripts of parliamentary speeches) and cross-lingual transfer from Spanish. Overall, our simple traditional SVM classifier performed best according to the shared task evaluation, ranking first in ideology detection and second in gender identification. We also demonstrate promising results for out-of-domain data and cross-lingual transfer learning.

Keywords

Political Ideology Detection, classification, SVM, BERT, XLM-Roberta, Transfer learning, Multitask learning

1. Introduction

Although difficult to define, a relatively uncontroversial definition of political ideology is 'a set of beliefs about the proper order of society, and how it can be achieved' [1, 2]. As the definition suggests, people's political ideologies, and especially those of politicians, have a significant impact on society. Much like other personal characteristics, such as gender, age, or native language [3, 4, 5], political ideology can help to understand individual and social behavior [6, 7, 8].

Besides analyzing social and political discourse, detecting ideology is a crucial step for correctly understanding political texts. Certain words or phrases have different intended meanings depending on the ideological position of speakers or authors [9, 10], and a message can also be understood differently based on the audience. In its extreme case, so-called 'dog whistles' [11] may allow politicians to send 'coded messages' to only a part of the public who share their ideological position. As a result, identifying political ideology is also important for natural language understanding tasks.

This paper describes our contribution to the EVALITA 2023 [12] shared task [13] on predicting political orientation and gender of politicians from social media posts in Italian. The task is defined as three related classification sub-tasks: (1) binary political orientation (`left`, `right`), (2) fine-grained political orientation (`left`, `moderate_left`, `moderate_right`, `right`) and (3) binary gender (`female`, `male`). We evaluated both simple linear classifiers and models based on deep pretrained networks. We also experimented with additional resources from different domains (parliamentary speeches) and similar datasets in another language (Spanish).

The rest of this paper reports on our approach and results. The code, the models hyper-parameters, and any additional resources used in our contribution can be found at https://github.com/fidan-c/PoliticIT23-ideology-detection.

2. Method and Experiments

2.1. Data

The PoliticIT dataset comprises between 80 and 100 tweets from 1751 politicians (1298 for training and development, 453 for testing). As usual, the test set was released only at the end of the competition. All tweet instances are anonymized by masking references to politicians, political parties and other Twitter account mentions. Further information on data collection and labeling can be found in Russo et al. [13].

Besides the PoliticIT dataset, we experiment with a few additional resources: transcripts of the parliamentary speeches from the Italian section of the parliamentary corpora collection ParlaMint [14] and the PoliticES 2023 shared task dataset [15, 16].

Parliamentary speeches We use the Italian section of the ParlaMint 3.0 pre-release, which contains speeches both from the Senate and the Chamber of Deputies, spanning from March 2013 to September 2022. We filter out

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT

ccoltekin@sfs.uni-tuebingen.de (Ç. Çöltekin); matteo.brivio@student.uni-tuebingen.de (M. Brivio); fidan.can@student.uni-tuebingen.de (F. Can)
https://coltekin.net/cagri/ (Ç. Çöltekin)
0000-0003-1031-6327 (Ç. Çöltekin); 0000-0001-6273-6900 (M. Brivio); 0009-0006-8900-2038 (F. Can)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper10.pdf

samples that belong to the chairperson and those that are less than 50 characters long. We also exclude speakers without a known party affiliation, as well as those who are affiliated with parties whose political orientation is either not specified or reported to be 'center' or 'big tent'.[1] Compared to the dataset released for the present shared task, the political orientation in the ParlaMint data is more fine-grained. For binary class labels, we map ParlaMint orientation labels to `left` if the class includes `L` (left) and to `right` if it includes `R` (right). We follow the same approach for the multi-class classification task, but mark orientation labels as `moderate` if they include `C` (center). For example, samples with original labels `CL` (center-left) or `CCL` (center to center-left) are mapped to `moderate_left`, while samples marked as `R` or `FR` (far-right) are mapped to `right`. The resulting corpus includes 167 `moderate_left`, 153 `moderate_right`, 89 `right` instances, and no instances with (non-center) `left` labels.

Compared to social media posts, parliamentary speeches tend to be much longer, and a few speakers have many more speeches than others in the corpus. To get a balanced dataset, we randomly select at most 10 speeches from each speaker and concatenate them with a separator token. The resulting corpus contains samples with an average of 2575.95 space-separated tokens per instance.

PoliticES 2023 shared task data The dataset consists of 2797 instances, each containing 80 tweets written by different users who share the same gender and political views. All tweets are anonymized, mirroring the anonymization procedure for the PoliticIT dataset. We use the PoliticES 2023 shared task data to conduct cross-lingual transfer learning experiments between Italian and Spanish. Further information about the data collection and labeling can be obtained from García-Díaz et al. [16].

2.2. Linear models

As a first, simple approach, we use linear support vector machines (SVMs) with bag-of-n-grams features. For all the results reported here, we train a separate linear SVM classifier for each task, with both word and character n-grams as features and weigh each feature using tf-idf. For each task, we tune the SVM margin parameter, `C`, (range 0.001-5.0), the maximum word n-grams (range 0-4) and the character n-grams (range 0-6). Lastly, we test whether n-grams should be case normalized or not (word, character, both or none). We draw 10 000 configurations uniformly from this hyper-parameter space and pick the setting with the highest average F1-score using 10-fold cross validation. The input to the linear models are all tweets belonging to each instance combined with an arbitrary separator symbol inserted between each one of them. To optimize the models trained on the ParlaMint data, we use the complete PoliticIT training set as the development set. Besides the predictions from the best model after the hyper-parameter search, we also experiment with the majority vote of the top-n best performing hyper-parameters. As the voting ensembles performed worse than the single-best model, we do not report their results here.

Since the binary ideology classifier was more accurate than the multi-class one in our initial experiments, we also include a post-processing step, where we adjust multi-class labels when they conflict with binary labels. If the multi-class classifier predicts (`moderate_`)`left` while the binary classifier predicts `right`, we set the multi-class label to `moderate_right`. Similarly, we resolve conflicts where the binary classifier predicts `left` by setting the multi-class label to `moderate_left`. All linear models are implemented using scikit-learn [17].

2.3. Transformer-based models

We also experiment with multi-task and cross-lingual transfer learning by fine-tuning pretrained language models. Specifically, we work both with a multilingual model, XLM-RoBERTa [18], and a monolingual version of BERT [19].[2] We customize both models, adding a 'common block' followed by three linear classification heads. Both solutions are implemented with PyTorch [20] and the Transformers library [21].

The 'common block' is a 2-layer, bi-directional Recurrent Neural Network which takes as input the language model's representation of the `CLS` token for each tweet in a given group. We include a ReLU activation function and a dropout layer after each RNN layer, except the last one. The model leverages the hard-shared parameters in the RNN and tries to learn the three tasks simultaneously (i.e. in a multi-task fashion) by minimizing the sum of their training losses: binary cross entropy loss for the two binary tasks and categorical cross entropy loss for the multi-class one. As observed by Liang and Zhang, it is reasonable to assume that in such a setting, the tasks have different levels of difficulty, so we rely on their *balanced multi-task learning* (BMTL) framework to mitigate this issue [22]. This requires wrapping each loss ℓ into a function h, such that $h(\ell) = \exp(\frac{\ell}{T})$, where T is a temperature parameter that needs to be tuned.

We train the XLM-R model in three different settings: first on the combination of the full PoliticES 2023 dataset and the PoliticIT 2023 training set, then on the Italian dataset alone and lastly only on the Spanish data. Fi-

[1] Political parties encouraging a broad spectrum of views among their members.

[2] https://huggingface.co/dbmdz/bert-base-italian-cased.

nally, the BERT-based model is also trained on the Italian dataset alone.

In an attempt to improve the multi-class ideology task score, we include a post-processing step similar to the one described in 2.2. If the binary score is `left` and the multi-class prediction is `moderate_right`, we set the multi-class label to `moderate_left` and the other way around if the two scores are `right` and `moderate_left`. Similarly, if the binary score is `left` and the multi-class prediction is `right`, we set the latter to `left` and the other way around if the two scores are `right` and `left`.

All models use the AdamW optimizer and a scheduler with a learning rate that decreases following the values of the cosine function. As a pre-processing step, we remove all punctuation marks and masking tokens originally used to anonymize the tweets (e.g., `[POLITICIAN]`, `[POLITICAL_PARTY]`), expand all hashtags and convert emojis into their text descriptions.

All models are trained for 15 epochs and their hyperparameters are optimized with Ray Tune [23], relying on a Bayesian optimization strategy.

3. Results

Our best performing system (based on linear SVMs) ranked first among 7 participating teams on both the binary and the multi-class ideology prediction tasks, achieving a macro-averaged F1-score of 92.82 and 75.15, respectively. With respect to the gender prediction task, our system ranked second with a score of 79.25.

Besides these simple models, we experimented with parliamentary speeches as an out-of-domain dataset and cross-lingual transfer, using pretrained language models. We summarize our results in Table 1.

Out-of-domain data To investigate the effectiveness of training on out-of-domain data, we train an SVM model on parliamentary speeches from the ParlaMint dataset and test it on the PoliticIT test-set. We present the results in Table 1 (indicated as SVM-ParlaMint).

ParlaMint includes both political orientation and gender data for the speakers in the Italian parliament. However, as noted in Section 2.1, it does not include speakers belonging to any (non-center) left parties for the multi-class classification. As a result, the scores for multi-class ideology detection are rather poor. Nonetheless, both the binary ideology and gender scores are not too far behind the models trained on the in-domain data.

Cross-lingual transfer The results obtained with monolingual and cross-lingual pretrained models, Italian BERT and XLM-RoBERTa, are also presented in Table 1. Contrary to our expectations, the BERT-based model performs rather poorly compared to SVM-PoliticIT, with the sole exception of the gender classification task, where the monolingual model achieves a macro-averaged F1-score of 81.17, outperforming the SVM model.

Even though it does not completely exceed the performance of the SVM-PoliticIT classifier, the multilingual model trained on both the Italian and the Spanish data improves the scores substantially. Specifically, the model achieves a macro-averaged F1-score of 80.56 in the gender task, outperforming SVMs. The effectiveness of cross-lingual transfer learning can be observed by examining the results of the monolingual BERT model which, with the only exception of gender, are significantly lower. However, it should be noted that XLM-RoBERTa and BERT are pretrained using different training regimes. To get a clearer understanding of the potential benefits of cross-lingual transfer, we trained XLM-RoBERTa on the Italian dataset alone. The scores we report for the latter model are marginally lower than those obtained with the model trained on both the Italian and the Spanish data, suggesting a beneficial effect of cross-lingual transfer learning. This is further corroborated by the results in a zero-shot setting, where the XLM-RoBERTa model was trained on Spanish data only. Although the scores are well below those of the other models, they are clearly better than random, indicating signal in the cross-lingual data.

4. Discussion and Concluding Remarks

We presented our experiments on how to identify political orientation and gender from social media posts as part of the EVALITA 2023 PoliticIT shared task. Our systems based on linear SVMs achieved the best overall scores. We also experimented with classifiers based on deep pretrained language models, focusing particularly on cross-lingual transfer.

Our first finding is that 'traditional' linear models show a better overall performance than the deep learning ones. Given their simplicity and lack of information other than immediate surface features in the training data, their success may come as a surprise. However, this is not the first time that such models have been found to perform similarly or better than deep networks [24, 25, 26, 27, just to name a few]. Similarly, the logistic regression classifiers of Mosquera [28] obtained the third place, only slightly behind the first two systems, in the PoliticES 2022 shared task. This, of course, may be due to a lack of proper tuning of the pretrained models. However, the fact that the linear SVMs outperformed all other participating teams (presumably the majority of them using pretrained models) shows that these simpler models have some strong merits in this task.

Our preliminary experiments on out-of-domain data

Model	Task	Precision	Recall	F1-score
SVM-PoliticIT	ideology binary	93.25	92.58	92.82
	ideology multi	80.12	73.41	75.15
	gender	81.72	77.71	79.25
SVM-ParlaMint	ideology binary	79.49	78.55	78.79
	ideology multi	44.89	56.55	49.49
	gender	73.39	76.73	74.17
BERT-based model	ideology binary	84.62	78.06	78.34
	ideology multi	73.72	63.09	64.12
	gender	80.92	81.43	81.17
XLM-R-based model (it)	ideology binary	89.75	86.55	87.14
	ideology multi	74.73	68.84	69.76
	gender	75.43	79.26	76.26
XLM-R-based model (es)	ideology binary	60.16	55.90	52.30
	ideology multi	37.42	23.53	17.40
	gender	58.12	53.12	33.04
XLM-R-based model (es-it)	ideology binary	90.16	90.28	90.21
	ideology multi	75.28	68.44	70.08
	gender	79.47	82.62	80.56

Table 1
Precision, recall and F1-scores of all models. All scores are macro-averaged scores on the official PoliticIT shared task test set.

show that training an SVM model on parliamentary speeches results in lower, but comparable performance to training on in-domain-data. The results clearly indicate that there is a considerable cross-domain signal for both ideology detection and gender prediction. Although we did not experiment with the use of both in- and out-of-domain data together in this study, the results highlight the potential for using out-of-domain data to improve ideology prediction in social media.

Our experiments also show promising results on cross-lingual transfer. The best results we achieve with pre-trained language models come from fine-tuning a cross-lingual model using both Spanish and Italian data. Part of this success is probably due to the similar methodologies followed by the organizers of both shared tasks. Further investigation of cross-lingual and cross-country transfer on ideology detection may shed light into universal and culture-specific aspects of 'ideology'.

Our best-performing model is a bag-of-words SVM classifier, relying only on token unigrams as features. Compared to the transformer-based ones, this model allows for better interpretability[3] with its unigram features revealing some linguistic indications of political orientation. Words such as *immigrazione* 'immigration', *clandestini* 'illegal immigrants' and *confini* 'borders' appear constantly in right-wing discourse, while terms such as *diritti* 'rights', *democrazia* 'democracy' and *solidarietà* 'solidarity' seem to point to more leftist views. The model also picked up unmasked tokens left in the data that could suggest a party or politician's name. This is the case with words like *Giorgia*, *FI* and *Democratico*, which allude to the current right-wing Prime Minister Giorgia Meloni, a well-known Italian right-wing party *Forza Italia*, and a prominent Italian left-wing party *Partito Democratico*, respectively. Further and better approaches to explain model behavior is another direction for future research that may shed light in narratives of different ideological groups.

[3]Although one should be cautious because tokens are not independent features, the weights assigned to individual tokens are, nevertheless, easier to interpret than weights of a neural network or weights in a model with overlapping features.

References

[1] J. T. Jost, C. M. Federico, J. L. Napier, Political ideology: Its structure, functions, and elective affinities, Annual review of psychology 60 (2009) 307–337.

[2] R. S. Erikson, K. L. Tedin, American public opinion: Its origins, content and impact, Routledge, 2015.

[3] C. Stachl, F. Pargent, S. Hilbert, G. M. Harari, R. Schoedel, S. Vaid, S. D. Gosling, M. Bühner, Personality research and assessment in the era of machine learning, European Journal of Personality 34 (2020) 613–631.

[4] H. Christian, D. Suhartono, A. Chowanda, K. Z. Zamli, Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging, Journal of Big Data 8 (2021) 1–20.

[5] F. Rangel, P. Rosso, M. Montes-y Gómez, M. Potthast, B. Stein, Overview of the 6th author profiling task at PAN 2018: multimodal gender identification

in Twitter, Working notes papers of the CLEF (2018) 1–38.

[6] M. Fatke, Personality traits and political ideology: A first global assessment, Political Psychology 38 (2017) 881–899. URL: http://www.jstor.org/stable/45095184.

[7] B. Verhulst, L. J. Eaves, P. K. Hatemi, Correlation not causation: The relationship between personality traits and political ideologies, American journal of political science 56 (2012) 34–51.

[8] J. A. García-Díaz, R. Colomo-Palacios, R. Valencia-García, Psychographic traits identification based on political ideology: An author analysis study on Spanish politicians' tweets posted in 2020, Future Generation Computer Systems 130 (2022) 59–74. URL: https://www.sciencedirect.com/science/article/pii/S0167739X21004921. doi:https://doi.org/10.1016/j.future.2021.12.011.

[9] G. Lakoff, Moral politics: How liberals and conservatives think, University of Chicago Press, 2010.

[10] M. Iyyer, P. Enns, J. Boyd-Graber, P. Resnik, Political ideology detection using recursive neural networks, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Baltimore, Maryland, 2014, pp. 1113–1122. URL: https://aclanthology.org/P14-1105. doi:10.3115/v1/P14-1105.

[11] W. Safire, Safire's political dictionary, Oxford University Press, 2008.

[12] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for Italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org <http://CEUR.org>, Parma, Italy, 2023.

[13] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian (2023).

[14] T. Erjavec, M. Ogrodniczuk, P. Osenova, N. Ljubešić, K. Simov, A. Pančur, M. Rudolf, M. Kopp, S. Barkarson, S. Steingrímsson, Ç. Çöltekin, J. de Does, K. Depuydt, T. Agnoloni, G. Venturi, M. Calzada Pérez, L. D. de Macedo, C. Navarretta, G. Luxardo, M. Coole, P. Rayson, V. Morkevičius, T. Krilavičius, R. Darǵis, O. Ring, R. van Heusden, M. Marx, D. Fišer, The ParlaMint corpora of parliamentary proceedings, Language resources and evaluation 57 (2022) 415–448. doi:10.1007/s10579-021-09574-0.

[15] J. A. García-Díaz, S. M. Jiménez Zafra, M. T. Martín Valdivia, F. García-Sánchez, L. A. Ureña López, R. Valencia García, Overview of PoliticEs 2022: Spanish author profiling for political ideology, Procesamiento del Lenguaje Natural 69 (2022) 265–272. doi:10.26342/2022-69-23.

[16] J. A. García-Díaz, S. M. Jiménez Zafra, M. T. Martín Valdivia, F. García-Sánchez, L. A. Ureña López, R. Valencia García, Overview of PoliticES2023@IberLEF: Political ideology detection in Spanish texts, Procesamiento del Lenguaje Natural 71 (2023).

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[18] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, CoRR abs/1911.02116 (2019). URL: http://arxiv.org/abs/1911.02116. arXiv:1911.02116.

[19] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186.

[20] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: An Imperative Style, High-Performance Deep Learning Library, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[21] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https:

//www.aclweb.org/anthology/2020.emnlp-demos.6.
[22] S. Liang, Y. Zhang, A simple general approach to balance task difficulty in multi-task learning, ArXiv abs/2002.04792 (2020).
[23] R. Liaw, E. Liang, R. Nishihara, P. Moritz, J. E. Gonzalez, I. Stoica, Tune: A research platform for distributed model selection and training, arXiv preprint arXiv:1807.05118 (2018).
[24] M. Medvedeva, M. Kroon, B. Plank, When sparse traditional models outperform dense neural networks: the curious case of discriminating between similar languages, in: Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial), Association for Computational Linguistics, Valencia, Spain, 2017, pp. 156–163. URL: https://aclanthology.org/W17-1219. doi:10.18653/v1/W17-1219.
[25] Ç. Çöltekin, T. Rama, Tübingen-Oslo at SemEval-2018 task 2: SVMs perform better than RNNs in emoji prediction, in: Proceedings of the 12th International Workshop on Semantic Evaluation, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 34–38. URL: https://aclanthology.org/S18-1004. doi:10.18653/v1/S18-1004.
[26] A. Caines, P. Buttery, REPROLANG 2020: Automatic proficiency scoring of Czech, English, German, Italian, and Spanish learner essays, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 5614–5623. URL: https://aclanthology.org/2020.lrec-1.689.
[27] K. Amponsah-Kaakyire, D. Pylypenko, J. Genabith, C. España-Bonet, Explaining translationese: why are neural classifiers better and what do they learn?, in: Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 2022, pp. 281–296. URL: https://aclanthology.org/2022.blackboxnlp-1.23.
[28] A. Mosquera, Alejandro mosquera at PoliticEs 2022: Towards robust Spanish author profiling and lessons learned from adversarial attacks, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022). CEUR Workshop Proceedings, CEUR-WS, A Coruna, Spain. D. Moctezuma, and V. Muniz-Sánchez, 2022.

URJC-Team at PoliticIT: Political Ideology Detection in Italian Texts Using Transformers Architectures*

Miguel Ángel Rodríguez-García[1,*,†]

[1]Universidad Rey Juan Carlos, Spain

Abstract

Over the years, psychologists have examined human personalities to understand their behaviours. This analysis has demonstrated the existence of several factors that are highly correlated to their conduct that supply clues about thought patterns. Consequently, in recent years the automatic prediction of personality traits has received considerable attention from the community. Thus, EVALITA proposes an evaluation campaign of Natural Language Processing, which suggests the PoliticIT task, which aims to recognise Twitter users' political polarity by analysing their comments. In particular, this task is composed of two subtasks focused on binary and multiclass classification problems. In this work, it is proposed a system which utilises Deep Learning architectures to address these subtasks. Three different pre-trained versions of the BERT transformer model are employed for each task. The outcomes of each generated model reached acceptable scores on the binary classification problems, but its performance dropped slightly on the multiclassification problem. In binary classification, 0.77 and 0.72 were achieved in gender and ideology tasks, respectively, and 0.56 in the multi-class.

Keywords

Natural Language Processing, Transformers architecture, Political Ideology Detection, Deep Learning,

1. Introduction

Personality is defined through people's behaviour, motivation, emotion and features of their thought patterns [1]. Besides, it accompanies us, impacting our daily lives significantly and affecting our decisions, satisfaction with life, well-being, happiness, preferences and desires [2]. The capability to automatically infer personality traits has an incredible amount of practical applications since it helps us understand human nature deeply, enabling us to discover how humans behave and think in determined situations [3].

Social Media has revolutionized the way that people interact, providing an open and direct path for communicating with each other anytime and anywhere and for other varied purposes like enjoyment or simply information access [4, 5, 6]. Hence, with the increasing popularity of this interactive way, Social Media has become a convenient resource for studying deeply human behaviour, considering personality traits and linguistic behaviour [7]. In this context, the target of this work was beyond this analysis, trying to understand the connection between the conduct of linguistic humans and their political ideology polarity.

This article describes the approach submitted to the challenge proposed in the EVALITA campaign for recognizing the political polarity in Italian texts [8, 9, 10]. It relates a system based on Deep Learning architectures like Encoders-Decoders and their families of masked-language models. Specifically, the proposed solution for the challenge is based on Transformers, where three different Bidirectional Encoder Representations from Transformers (BERT) are employed on each subtask. The selection was taken as a consequence of a literature review, where it was studied the highest performance of these models on natural language classification problems.

The rest of the paper is organised as follows. Section 2 a brief overview of the related work. Section 3 details the datasets' content delivered for each subtask and the proposed system's architecture. Section 4 analyses the results achieved on each subtask in the challenge. Finally, Section 5 compiles the findings gained facing this challenge.

2. Related work

Several studies discovered the party ideology in humans is strongly related to their personal traits [11, 12]. Its analysis is relevant since it provides more detailed information about how humans behave and thinks in determined situations [3]. Hence, hereafter, it is provided with a cutting-edge context about the proposed solutions in political ideology recognition. The study begins with the system proposed by Baly et al., in [13], where they employed two models based on two deep learning architectures: Long Short-Term Memory networks (LSTMs) and Bidirectional Encoder Representations from Transformers (BERT) to predict the political polarity of news arti-

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT

*Corresponding author.

†These authors contributed equally.

miguel.rodriguez@urjc.es (M. Rodríguez-García)

0000-0001-6244-6532 (M. Rodríguez-García)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper11.pdf

cles. Then, in the experiments, they fine-tuned the hyperparameters of both models, playing with the length of the inputs, sizes of the architecture, learning rates, and batch sizes, among others. As a result, they concluded that the BERT overperformed the LSTM architecture in the experiments accomplished. In the same way, Iyyer et al., in [14] applied a Recursive Neural Network (RNN) model for identifying political polarity signals at the sentence level. For the experiments, they created a new dataset that combines Convote, an initial dataset created, and a filtered version of the Ideological Books Corpus (IBC), a collection of resources whose authors had well-known political leanings. In the results, they created a baseline by selecting Machine Learning techniques like Logistic regression to compare with more complex strategies based on Deep Learning like RNN. The results show how RNN clearly outperform traditional methods. Finally, in [15], Ozturk and Ozcan studied the application of varied Machine Learning Strategies such as Transformers, Long Short Term Memory (LSTM), and Support Vector Machines, among others, on the ideology classification problem. After testing different settings, they closed that a couple of Transformers architectures employed in the task achieved greater precision than others. Therefore, given the notable performance of the Transformers architectures, it was decided to select this choice to deal with the subtasks proposed.

3. Material and methods

This section describes the methods developed to face the tasks proposed and the distribution of the datasets delivered.

3.1. Data

The dataset was composed of tweets harvested from Twitter user accounts, whose political affiliation could be easily deduced since they have a clear link to the party they belong to [16, 17]. The collected tweets were grouped into clusters to avoid ethical and privacy issues and preprocessed to anonymize users' names and remove those tweets, including web content, news, and links. Furthermore, they were labelled, considering gender: male and female, and political spectrum from two perspectives: i) binary: right and left, and ii) multiclass: right, moderate right, left, and moderate left. As a result, the dataset was composed of a set of clusters, which contained 80 tweets each. The delivered dataset was organized into 80%-20% for training and testing in the proposed challenge. Table 1 depicts the distribution of the datasets delivered for practice and evaluation.

The distribution analysis shows the imbalance of the dataset on each task. The highest is on the practice dataset, in which the label `'left'` only has 11 clusters against the label `'right'`, which has 37. For its part, in the dataset delivered for evaluation, the less imbalanced dataset is in `'gender'`, where labels `'male'` and `'female'` differ in 2 clusters. On the contrary, the more noticeable difference is again on `'ideology_multiclass'`, where there is decompensation between the four types of labels included.

3.2. Method

The system proposed for the challenge is composed of a set of three modules: i) the cleaner, which is responsible for removing useless information from tweets; ii) the classifier, which represents the main module in the system since it carries out categorization tasks; iii) the evaluator, which aims at quantifying the performance of the system. The classifier is composed of three different versions of the BERT model [18]. Each one was pre-trained differently by using dissimilar datasets. The version dedicated to the subtask gender was trained by employing text obtained from Wikipedia dump and OPUS corpora collection. [1] The model responsible for identifying the political polarity in the binary classification problem was the XXL version of before, which was trained by employing the OSCAR corpus. [2] Finally, for the multiclassification subtask, it was utilized AlBERTo, a language model trained to understand the users' jargon in social networks [19]. [3] Figure 1 depicts the modular architecture proposed.

The system' architecture has been configured as a pipeline, and it works as follows: two inputs are defined to receive training and testing datasets. When the training dataset is provided, the cleaner module removes the elements in the tweets that do not provide valuable information as links, symbols, and emojis. Then, depending on the faced classification task, a different model is trained in the classifier module. Next, the evaluator employes the test dataset to quantify the precision of the model. As a result, it provides three different outcomes, the resulting CSV file, the main classification metrics and the confusion matrix.

4. Results and Discussion

This section details the outcomes obtained on each subtask delivered in the challenge. The metrics selected to assess the performance of the proposed systems in the challenge are precision, recall, and F1-score. Table 2 scrutinizes the results achieved by the proposed architecture on each classification task.

[1] https://huggingface.co/dbmdz/bert-base-italian-cased
[2] https://huggingface.co/dbmdz/bert-base-italian-xxl-cased
[3] https://huggingface.co/m-polignano-uniba/bert_uncased_L-12_H-768_A-12_italian_alb3rt0

		Practise		Evaluation	
		Train	Test	Train	Test
gender	male	23	23	810	318
	female	25	25	488	135
Total		48	48	1298	453
ideology_binary	right	37	37	578	205
	left	11	11	720	248
Total		48	48	1298	453
ideology_multiclass	right	10	10	131	51
	moderate_right	27	27	447	154
	left	1	1	558	148
	moderate_left	10	10	162	100
Total		48	48	1298	453

Table 1
The distribution of the delivered dataset in clusters.

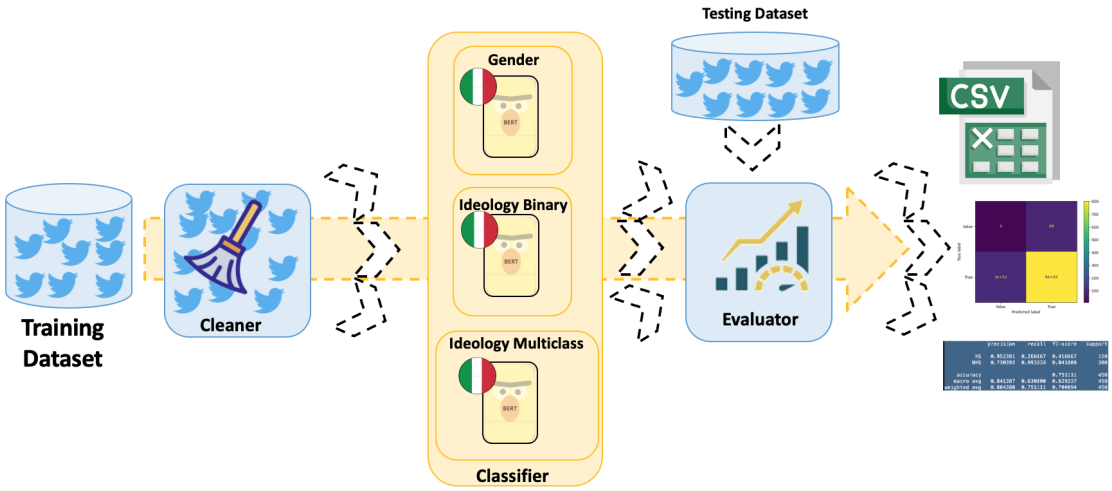
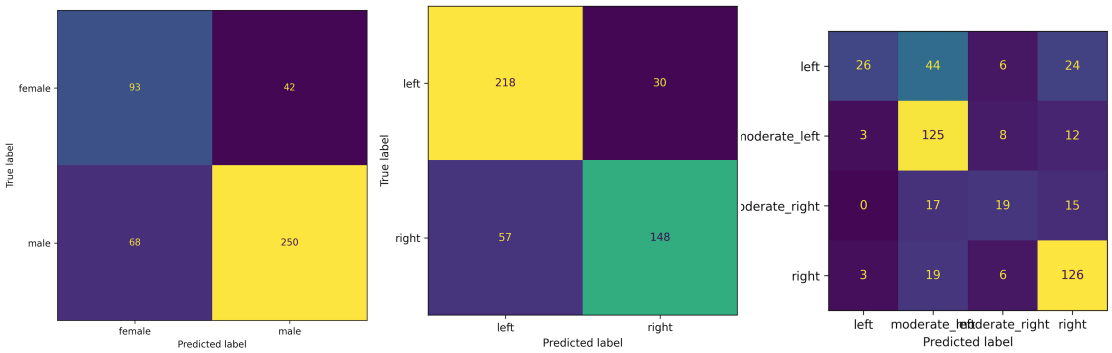


Figure 1: Architecture of the proposed system.

		Evaluation		
Task	Label	Precision	Recall	F1-score
gender	female	0.51	0.63	0.56
	male	0.78	0.72	0.75
	Macro AVG	0.65	0.68	0.66
ideology_binary	right	0.76	0.85	0.8
	left	0.8	0.69	0.74
	Macro AVG	0.78	0.77	0.77
ideology_multiclass	right	0.75	0.87	0.81
	moderate_right	0.54	0.42	0.47
	left	0.85	0.31	0.45
	moderate_left	0.65	0.89	0.75
	Macro AVG	0.7	0.62	0.62

Table 2
Results on the official test set.



(a) Gender Task. (b) Ideology Binary Task. (c) Ideology Multiclass Task.

Figure 2: Confusion matrix computed on each task.

The table was divided into three zones, one per each subtask delivered. The best results were achieved in classifying the political polarity, primarily on tagging tweets as a `'right'` in binary and multiclass classification problems, where both employed models obtained an F1-score of 0.8 and 0.81, respectively. If we look at the dataset distribution in Table 1, both cases did not contain the highest number of samples. Therefore, from this behaviour, it can be deduced that, even when the dataset is non-balanced and the amount of training data is not abundant, both models could recognise the correct label with a high percentage. Moreover, these high results suggest that the samples found in the dataset and used during the training phase contained more discriminate features, improving the models' discrimination capacity. Conversely, the low results obtained on `'ideology_multiclass'`, where the precision dropped to 0.54 and 0.65 in `'moderate_right'` and `'moderate_left'`, respectively, reveals that despite the highest number of training data, the outcomes were not extremely high, indicating the characteristics extracted from tweets were not discriminant enough to teach the model how to recognise these labels during the classification tasks. Finally, to have a deeper understanding of the models' behaviour, Figure 2 shows the confusion matrices of each classification task.

Starting for the gender task, the confusion matrix reveals that the model had great difficulties in characterizing tweets written by males since it misclassified 68 against 42 mistakes made in female classification. For the political polarity identification, in the binary classification problem, the models made more mistakes in recognizing right than left. In the multilabelling task, the errors were in the label `'left'` when the model tried to differentiate this against `'moderate_left'` and `'right'`, making 44 and 24 mistakes, respectively. From this amount of elevated errors, it can be deduced that the text annotated by these labels, `'right'` and `'moderate_left'`, contains similar features to `'left'` since the model can not differ between them. Consequently, a more exhaustive study of the dataset is required to identify precisely the most discriminant features between these labels and reduce the ratio of mistakes made.

At the close of this section, it is compared the results obtained by the architecture proposed against the three first participants and the baseline delivered. Table 3 shows an excerpt of the official leader board, which depicts the F1-score obtained by these proposals.

With regard to the baseline, which combines the Bag of Words (BoW) and the logistic regression, it achieves better results than our proposal on the binary task of ideology, reaching 0.81 against 0.77. However, it is obtained much better results in the other tasks, doubling the performance in the multi-classification task. This imbalance between the two approaches reveals the influence of the choice of features on the behaviour of the techniques since it is compared two ways of extracting features, a set of isolated words versus embeddings, in other words, words versus semantics. The higher results obtained by the baseline show that the selected features are highly discriminative and influence the highest results. The lower results show exactly the opposite. There are similar features assigned to these classification tags that cause the approach to make mistakes, which has a negative impact on the results obtained. On the other hand, the results obtained by the first three proposals in the ranking are very far from the results obtained by the other proposals. Only in the multiclassification task, the proposed architecture can approximate the performance of the third classified opponent, which means there is still work to be done in fine-tuning the architecture and extracting features.

Team Name	F1Gender	F1Ideology Binary	F1Ideology Multiclass
TuebingenPoliticIT	0.79	0.93	0.75
INFOTEC-LaBD	0.82	0.86	0.72
extremITA	0.77	0.92	0.62
...	...	...	...
NLP_URJC	0.66	0.77	0.62
UMU_TEAM	0.53	0.81	0.36

Table 3
The official leader board.

5. Conclusions

This article describes the proposed system for EVALITA, an evaluation campaign focused on Natural Language Processing and Speech challenges for the Italian language in 2023, promoted by the Italian Association of Computational Linguistics (AILC). In particular, this work faced the PoliticIT challenge, which motivates the development of smart systems capable of recognizing the ideological polarity in italian texts. The challenge included three subtasks that address three different classification problems, two binary, where it is demanded to categorize gender and political polarity and one multiclass, where the political polarity was extended from two to four different classes. To deal with these tasks, a system based on the Transformers models. The system includes three versions of the BERT model pre-trained using three datasets. The best performance obtained by the system was on the binary political polarity classification subtask. Not too far are the remaining tasks, in which the system gains relatively similar scores. In spite of the results obtained being quite elevated, there is still a significant range of improvement.

In future work, several lines would be interesting to explore for extending the architecture presented here. Firstly, it would be interesting to use augmentation techniques to study their effects on the Transformers models selected. A more detailed dataset study would be interesting to address, primarily to identify discriminant features that can boost the system's performance.

Acknowledgments

This research has been partially supported by grants: PID2021-125709OA-C22, funded by MCIN/AEI/10.13039/501100011033, and "ERDF A way of making Europe"; P2018/TCS-4566, funded by Comunidad de Madrid and European Regional Development Fund, and "Programa para la Recualificación del Sistema Universitario Español 2021-2023".

References

[1] Y. Mehta, N. Majumder, A. Gelbukh, E. Cambria, Recent trends in deep learning based personality detection, Artificial Intelligence Review 53 (2020) 2313–2339.

[2] G. Durand, J. Lobbestael, The relationship between psychopathic personality, well-being, and adaptive traits in undergraduates, Personality and Individual Differences 204 (2023) 112050.

[3] P. Sharma, D. Jain, A. Dhull, Psychology approach: Through ai, ml and dl, Psychology 7 (2021).

[4] G. Appel, L. Grewal, R. Hadi, A. T. Stephen, The future of social media in marketing, Journal of the Academy of Marketing science 48 (2020) 79–95.

[5] F. Monti, F. Frasca, D. Eynard, D. Mannion, M. M. Bronstein, Fake news detection on social media using geometric deep learning, arXiv preprint arXiv:1902.06673 (2019).

[6] L. Xiao, J. Mou, Social media fatigue-technological antecedents and the moderating roles of personality traits: The case of wechat, Computers in Human Behavior 101 (2019) 297–310.

[7] G. Mavis, I. H. Toroslu, P. Karagoz, Personality analysis using classification on turkish tweets, International Journal of Cognitive Informatics and Natural Intelligence (IJCINI) 15 (2021) 1–18.

[8] D. Russo, S. M. Jiménez-Zafra, J. A. García-Díaz, T. Caselli, M. Guerini, L. A. Ureña-López, R. Valencia-García, Overview of PoliticIT2023@EVALITA: Political Ideology Detection in Italian Texts, 8th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2023) (2023).

[9] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[10] D. Russo, S. M. Jiménez Zafra, J. A. García-Díaz, T. Caselli, M. Guerini, L. A. Ureña-López, R. Valencia-García, PoliticIT at EVALITA 2023: Overview of the Political Ideology Detection in Italian Texts Task, in: M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi (Eds.), Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop, 2023.

[11] B. N. Bakker, Y. Lelkes, A. Malka, Reconsidering the link between self-reported personality traits and political preferences, American Political Science Review 115 (2021) 1482–1498.

[12] A. Furnham, M. Fenton-O'Creevy, Personality and political orientation, Personality and Individual Differences 129 (2018) 88–91.

[13] R. Baly, G. D. S. Martino, J. Glass, P. Nakov, We can detect your bias: Predicting the political ideology of news articles, arXiv preprint arXiv:2010.05338 (2020).

[14] M. Iyyer, P. Enns, J. Boyd-Graber, P. Resnik, Political ideology detection using recursive neural networks, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2014, pp. 1113–1122.

[15] O. Ozturk, A. Özcan, Ideology detection using transformer-based machine learning models (2022) 1–23. doi:10.13140/RG.2.2.12303.51362.

[16] J. A. García-Díaz, R. Colomo-Palacios, R. Valencia-García, Psychographic traits identification based on political ideology: An author analysis study on Spanish politicians' tweets posted in 2020, Future Generation Computer Systems 130 (2022) 59–74.

[17] J. A. García-Díaz, S. M. Jiménez-Zafra, M.-T. M. Valdivia, F. García-Sánchez, L. A. Ureña-López, R. Valencia-García, Overview of politices 2022: Spanish author profiling for political ideology, Procesamiento del Lenguaje Natural 69 (2022) 265–272.

[18] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[19] M. Polignano, P. Basile, M. De Gemmis, G. Semeraro, V. Basile, et al., Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets, in: CEUR Workshop Proceedings, volume 2481, CEUR, 2019, pp. 1–6.

GeoLingIt at EVALITA 2023: Overview of the Geolocation of Linguistic Variation in Italy Task

Alan Ramponi[1,*], Camilla Casula[1,2]

[1]Fondazione Bruno Kessler (FBK), Digital Humanities Unit – Trento, Italy
[2]University of Trento, Department of Information Engineering and Computer Science – Trento, Italy

Abstract

GeoLingIt is the first shared task on geolocation of linguistic variation in Italy from social media posts comprising content in language varieties other than standard Italian (i.e., regional Italian, and languages and dialects of Italy). The task is articulated into two subtasks of increasing complexity for which only textual content is allowed: i) *coarse-grained geolocation*, aiming at predicting the region in which the variety expressed in the post is spoken, and ii) *fine-grained geolocation*, aiming at predicting its exact coordinates. Both tasks can be either at the country level (*standard track*) or restricted to a linguistic area of choice (*special track*). GeoLingIt has attracted wide interest at the Evalita 2023 evaluation campaign with 37 registrations and 35 submitted runs. In this paper, we present the task and data, the evaluation criteria, the participants' results, an analysis of their approaches, and the main insights from the shared task.

Keywords

Natural language processing, computational sociolinguistics, linguistic variation, linguistic diversity

1. Introduction

Italy is characterized by an astonishing linguistic diversity that makes it a unique landscape in Europe [1]. Besides standard Italian, a large number of local languages, their dialects, and regional varieties of standard Italian (i.e., regional Italian) are spoken across the country [2]. While Italian is employed in all formal settings in its standard form, in informal situations it is natural to observe Italian speakers to use (even unwittingly) regional forms of Italian (e.g., *guaglione*, *toso*, and *caruso* for "young man", typically in Campania, Lombardy-Veneto, and Sicily areas, respectively), or to code-switch their local language varieties with the national language.

Local languages and their dialects evolved from Vulgar Latin like Italian, and they mostly have no established orthography insofar as they are primarily used in spoken settings. On the other hand, regional forms of Italian derive from a geographical differentiation of Italian due to influences by the former [3], are largely used in both oral and written informal contexts, and typically follow Italian spelling conventions. When it comes to user-generated texts on social media, which are informal and feature linguistic patterns from spoken language [4, 5], we observe that not only regional Italian is naturally present, but also local language varieties of Italy are employed, albeit at various degrees. This can be attributed to their rediscovery as "additional expressive resources" [6], especially by the youngest generations. User-generated texts comprising language varieties other than standard Italian open opportunities for the study of linguistic variation in Italy, and can ultimately help in enriching and complement linguistic atlases.

In this paper, we present GeoLingIt, the first shared task on geolocation of linguistic variation in Italy from social media posts from Twitter containing content other than standard Italian. GeoLingIt has been organized as part of the Evalita 2023 evaluation campaign [7], and relies on DiatopIt [8], a corpus of geolocated tweets exhibiting regional Italian use, code-switching between Italian and local language varieties, or fully written in the latter. Compared to previous geolocation shared tasks at international venues [9, 10, 11], GeoLingIt is focused on Italy and tailored to variation across language varieties, and it thus minimizes the effect of spurious, highly-localized lexical items (e.g., mentions of events, places, or tourist attractions) on prediction of linguistic areas. In the following, we present details on GeoLingIt, the results obtained by participant teams, and the main insights from the shared task.

2. Task description

The GeoLingIt shared task deals with the geolocation of linguistic variation in Italy from Twitter posts comprising content in language varieties other than standard Italian (i.e., regional Italian, and languages and dialects of Italy). It aims to advance the study of linguistic variation in Italy, provide means to complement qualitative-driven linguistic atlases, and sensitize the community on the rich linguistic landscape of the country.

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
*Corresponding author.
alramponi@fbk.eu (A. Ramponi); ccasula@fbk.eu (C. Casula)
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper12.pdf

2.1. Standard and special tracks

GeoLingIt is organized into two tracks. In the *standard track*, the focus of the task is at the country level (i.e., comprising all language varieties of Italy), whereas in the *special track*, the task is restricted to a linguistic area chosen by participants[1] (e.g., the *Gallo-Italic area*, including language varieties spoken in Piedmont, Lombardy, Liguria, and Emilia-Romagna regions) to favor the emergence of microvariation insights. For both tracks, two subtasks of increasing complexity are possible: *coarse-grained geolocation* (Section 2.2) and *fine-grained geolocation* (Section 2.3).

2.2. Subtask A: Coarse-grained geolocation

Given the text of a tweet exhibiting regional Italian features or (partially or fully) written in local languages and dialects of Italy, predict the administrative region in which the variety expressed in the post is spoken. This is a classification task, i.e., one among n regions of Italy has to be predicted. In the case of the *standard track*, this matches all regions of Italy[2] ($n = 20$), whereas in the *special track*, it corresponds to the subset of regions k of the linguistic area under consideration ($n = k$). This subtask is applicable for the *special track* if $k \geq 2$ regions are represented in the chosen area.

2.3. Subtask B: Fine-grained geolocation

Given the text of a tweet exhibiting regional Italian features or (partially or fully) written in local languages and dialects of Italy, predict the location, in terms of longitude and latitude coordinates, in which the variety expressed in the post is spoken. This is a double regression task, i.e., a pair of real-valued numbers has to be predicted. The difference between *standard* and *special* tracks is here the extent of the area being considered. This subtask overcomes the simplification of *coarse-grained geolocation* (Section 2.2), aiming to uncover fine-grained linguistic variation. Indeed, language varieties of Italy lie on a continuum and often cross administrative region borders.

3. Data

GeoLingIt is based on DiatopIt [8], a corpus of social media posts from Twitter specifically focused on language variation in Italy. All tweets have associated geolocation information and region labels, and have been sampled to contain either regional Italian usage or content in local language varieties of Italy. A multi-stage data collection process has been followed based on data-driven out-of-vocabulary tokens (from posts over a period of 2 years) which have been curated manually. Under-represented areas from the resulting posts have been then augmented by employing the lexical artifacts package [13]. The corpus consists of 15,039 posts from a 2-year time frame (from `2020-07-01` to `2022-06-30`) to minimize period-related biases. For more details, we refer the reader to Ramponi and Casula (2023) [8].

Data splits During the development stage, participant teams are provided with the original training and development splits of DiatopIt. These splits consist of 13,669 and 552 examples, respectively. While the training set comprises content from all over the country, the development set contains data from 13 out of 20 regions.[3] Teams are allowed to use alternative splits and even augment the dataset at their will, with the only constraint to not use external Twitter data since some tweets can be part of the test set. The (unlabeled) test set is then released during the evaluation window for allowing teams to submit their predictions, and comprises 818 examples from the same regions in the development set plus examples from $1 \leq j \leq 7$ additional regions unknown to participants during both development and evaluation stages. At the end of the evaluation window, the $j = 4$ additional regions in the test set have been communicated to participants.[4] Splits match the original data partitions of DiatopIt; we thus refer the reader to Ramponi and Casula (2023) [8] for details on statistics and distribution.

Data format The corpus splits are in the form of `tsv` files, i.e., a tab-separated format, with an example per line and the first line as header. Each example has `id` and `text` columns. For the *coarse-grained geolocation* subtask, data files additionally include a `region` column, whereas data files for the *fine-grained geolocation* subtask include `latitude` and `longitude` columns. As a result, the instances in both the subtasks are the same, and differ according to the label column(s). The content of such columns is described below:

- **id**: a unique identifier, different from the original tweet identifier to preserve user's anonymity;
- **text**: the text of the tweet, with anonymized user mentions, email addresses, URLs, and location strings deriving from cross-platform posting;

[1] Participants have been provided with the renowned linguistic map by Pellegrini (1977) [12] to encourage linguistically-grounded proposals, and requests have been approved based on motivation and relevance of the area from a linguistics perspective.

[2] These are: Abruzzo, Aosta Valley, Apulia, Basilicata, Calabria, Campania, Emilia-Romagna, Friuli-Venezia Giulia, Lazio, Liguria, Lombardy, Marche, Molise, Piedmont, Sardinia, Sicily, Tuscany, Trentino-Alto Adige, Umbria, and Veneto.

[3] Regions in the development set: Apulia, Calabria, Campania, Emilia-Romagna, Friuli-Venezia Giulia, Lazio, Liguria, Lombardy, Piedmont, Sardinia, Sicily, Tuscany, and Veneto.

[4] Regions in the test set: the regions in the development set plus Abruzzo, Marche, Trentino-Alto Adige, and Umbria.

Table 1
Examples from the DiatopIt corpus [8] used for the GeoLingIt shared task. Columns in *yellow* and *gray* are included in subtask A and B files, respectively. *Note: IDs have been changed and texts have been slightly redacted to preserve users' anonymity.*

id	text	region	latitude	longitude
1	chiov' tutt a jurnat', ce serv' o mbrell'	Campania	40.92589400	14.12023790
2	ho così sonno che me bala l'oeucc	Lombardia	45.46134530	9.15933655
3	da caruso anche io ci andavo spesso!	Sicilia	37.46007165	15.03084825

- **region**: the administrative region from where the tweet has been posted, in a string format (for *subtask A* only);
- **latitude**: a real-valued number representing the latitude coordinate from where the tweet has been posted (for *subtask B* only);
- **longitude**: a real-valued number representing the longitude coordinate from where the tweet has been posted (for *subtask B* only).

The `latitude` and `longitude` coordinates are computed by taking the central point from the 4-point bounding box of city areas as provided by the Twitter APIs. Note that coordinates do not correspond to specific places within cities, but instead represent cities as a whole (i.e., posts originated within the same city have the same coordinates). We provide examples from the corpus in Table 1.

4. Evaluation

During the evaluation phase, participant teams are allowed to submit up to 3 runs (i.e., predictions on the unlabeled test set) for each track and subtask. In all the setups, only textual content can be used. We here present the metrics used for assessing the performance of runs (Section 4.1) and the baselines we provide (Section 4.2).

4.1. Metrics

Due to the different nature of *coarse-grained geolocation* and *fine-grained geolocation*, we employ different evaluation metrics for the subtasks. Subtask-specific metrics are the same for both *standard* and *special* tracks.

Subtask A The submitted runs are evaluated using macro-averaged precision, recall, and F_1 score on the n regions of Italy under consideration. For the *standard track*, this matches all the administrative regions in the test set ($n = 17$, cf. Section 3, "Data splits"), whereas for the *special track*, this corresponds to the k regions in the chosen linguistic area that are also represented in the test set ($n = k$, cf. Section 2.2). Runs are ranked by macro F_1 score and presented in separate rankings (i.e., one for the *standard track*, and one for each chosen subset of administrative regions in the *special track*).

Subtask B Since a level of detail to the centimeter is unnecessary to the study of linguistic variation, runs are evaluated using the average distance in kilometers (km) of predicted coordinates from gold coordinates (the lower the better) on either the whole test set (for the *standard track*) or the subset representing the linguistic area chosen by participants (for the *special track*). We employ the Haversine formula as implemented in the `haversine` (v2.8.0) package[5] for computing distance. Runs are presented in separate rankings (i.e., one for the *standard track*, and one for each area in the *special track*) and ordered by increasing average distance in kilometers.

4.2. Baselines

We use the same baselines for both tracks. For subtask A, we provide a most frequent baseline and a logistic regression baseline. For subtask B, we provide a centroid baseline and a k-nearest neighbors baseline.

Most frequent A baseline that always guesses the most frequent administrative region in the training set (i.e., Lazio) for all test set instances.

Logistic regression A machine learning classifier with default `scikit-learn` (v1.2.2)[6] hyperparameters that employs count vectorizer with unigrams for feature extraction and operates on original text casing.

Centroid A baseline that computes the center point (in terms of latitude and longitude) from the training set and predicts it for all test instances.

k-nearest neighbors (kNN) A machine learning regressor with default `scikit-learn` hyperparameters that employs count vectorizer with unigrams for feature extraction and operates on original text casing.

5. Participants and results

A total of 35 runs have been submitted to the GeoLingIt shared task: 26 runs (6 teams) for the *standard track* and

[5] `haversine` package: https://github.com/mapado/haversine
[6] `scikit-learn` library: https://scikit-learn.org

9 runs (2 teams) for the *special track*. Specifically, for the *standard track* we received 14 runs (5 teams) for subtask A and 12 runs (5 teams) for subtask B, whereas for the *special track* 6 runs (2 teams) have been submitted for subtask A (i.e., *Tuscany-Lazio area* and *Gallo-Italic area*) and 3 runs (1 team) have been tailored at subtask B (i.e., *Gallo-Italic area*). Overall, GeoLingIt has been one of the most participated shared tasks at Evalita 2023 [7] and attracted interest of heterogeneously composed teams with up to 7 individuals, from master students to senior academic researchers.

5.1. Overview of participant teams

In the following, we provide a summary of the approaches employed by participant teams. We refer the reader to their description papers for additional details.[7]

baρtti [14] The team participated in both subtasks for the *standard track*. For subtask A, they experimented with multi-task learning, a transformer-based and logistic regression model ensemble, and contrastive pre-training of a BERT-based Italian model on augmented subtask data. Augmentation uses a vocabulary built from online sources to create examples by randomly substituting words with lexical items from varieties spoken in the same or different regions. For subtask B, they leveraged data from both subtasks in a multi-task setting using either a BERT-based Italian model or the model that underwent continuous pre-training in subtask A, also testing a rectification module to adjust predictions outside land to the closest point within Italy's boundaries.

DANTE [15] The team focused on further pre-training BERT-based Italian language models and participated in both subtask A and B for the *standard track*. Specifically, they experimented with two multi-task pre-training setups, namely task-specific learning and joint learning, with dialect and token classification objectives, using texts collected from external sources. Fine-tuning is then done in a single task setup on relevant subtask data. In both subtasks, they also proposed ensembles of their best-performing models.

extremITA [16] The team proposed two *one-for-all* models, designed to tackle all the challenges at Evalita 2023. The first model is based on the IT5 encoder-decoder architecture, whereas the second one is an instruction-tuned model built upon LLaMA. For fine-tuning, they used data from all Evalita 2023 challenges and encoded the tasks as prompts. The team submitted a run for model for both subtasks of the *standard track*.

galliz [17] The team proposed a hybrid approach for subtask A, and participated in both the *standard track* and *special track*. Specifically, they combined the predictions given by *i)* an English pre-trained BERT classifier, previously fine-tuned on augmented GeoLingIt training data, and *ii)* a dictionary-based algorithm derived from external lexical sources. They then tested different hyperparameter setups. As regards data augmentation, the team fine-tuned an Italian word embedding model on the training set, and leveraged word vector similarities to create new training examples by substituting a single word per post with a close word in the embedding space.

Salogni [18] The team tested different transformer-based models pre-trained on Italian texts, with a set of hyperparameter settings (e.g., hidden layers, activation functions). They submitted a single run for the *standard track*, subtask B, based on a UmBERTo language model.

SCG The team participated to both tracks and experimented with logistic regression and support vector machines for subtask A, and linear regression and kNN regression for subtask B.[8] They did not submit a report and we are thus unable to discuss further their approach.

5.2. Results

In this section, we summarize the results of participant teams in both subtask A and B for the *standard track* (Section 5.2.1) and the *special track* (Section 5.2.2).

5.2.1. Standard track

We present the results divided by subtask below.

Subtask A: Coarse-grained geolocation In Table 2, we report the results on the test set for all runs submitted by teams participating in subtask A, ranked by macro F_1.

All runs by the *DANTE* team obtained the best results in the subtask, with improvements ranging from 5.52 to 10.10 macro F_1 points compared to the best run by the team that ranked second (*galliz*). The best-performing system by *DANTE* (run 3) is an ensemble of transformer-based classifiers originally pre-trained on Italian texts, which have been further pre-trained in a multi-task fashion on external data from Dialettando[9] and Wikipedia editions for local language varieties of Italy with region-centric objectives. The best submission by *galliz* (run 1) is an equally-weighted ensemble of a dictionary-based algorithm (based on Dialettando and GeoLingIt) and an English BERT model fine-tuned on augmented subtask A data, whereas the best run for *baρtti* (run 2) relies on a

[7] Indeed, we do not include the specific model versions and hyperparameter choices of participants' systems due to space constraints.

[8] We thank the SCG team for providing us with this information.

[9] "Dialettando" website: https://www.dialettando.com

Table 2
Results on the test set for the *standard track*, subtask A. Baselines are italicized and highlighted in *yellow*.

	Team	Run	P	R	F_1
1	DANTE	(3)	79.46	63.75	66.30
2	DANTE	(2)	66.98	62.65	63.93
3	DANTE	(1)	65.18	60.09	61.72
4	galliz	(1)	82.94	52.25	56.20
5	baρtti	(2)	67.97	51.62	53.18
6	galliz	(3)	74.58	49.49	52.08
7	baρtti	(3)	52.93	51.75	51.74
8	baρtti	(1)	56.05	51.68	51.72
9	galliz	(2)	68.98	45.36	47.74
	Log. reg.		*62.19*	*42.43*	*46.11*
10	extremITA	(1)	72.14	38.84	39.99
11	extremITA	(2)	65.03	37.62	38.18
12	SCG	(2)	12.92	9.82	9.28
13	SCG	(1)	10.15	9.97	9.04
14	SCG	(3)	10.42	6.60	7.85
	Most freq.		*4.47*	*21.15*	*7.38*

Table 3
Results on the test set for the *standard track*, subtask B. Baselines are italicized and highlighted in *yellow*.

	Team	Run	Avg dist (km)
1	baρtti	(3)	97.74
2	baρtti	(1)	98.79
3	DANTE	(3)	110.35
4	DANTE	(2)	112.58
5	DANTE	(1)	114.00
6	baρtti	(2)	120.02
7	extremITA	(1)	126.10
8	Salogni	(1)	128.19
9	extremITA	(2)	145.15
	kNN		*263.35*
10	SCG	(1)	280.99
	Centroid		*281.04*
11	SCG	(2)	281.20
12	SCG	(3)	289.91

transformer-based classifier, pre-trained on Italian texts, that has been further pre-trained in a contrastive learning fashion with subtask A data, preemptively augmented with a word substitution approach based on a vocabulary derived from Dialettando and Wikipedia content. While all teams outperformed the most frequent baseline, all runs by *extremITA* and *SCG* teams achieved worse results than the logistic regression baseline.

From a closer look, we observe that F_1 scores obtained by participants' runs greatly differ across regions (Figure 1). Campania, Lazio, Sardinia, Sicily, and Veneto are the easiest to classify. As expected, Abruzzo, Marche, Trentino-Alto Adige, and Umbria are instead among the regions with the lowest scores on average. This is mainly because posts from those regions have been excluded on purpose from the development set, and only few tweets are available in the training set, making traditional learning and tuning challenging. As a result, most instances from those regions are typically classified as neighboring regions in which similar varieties are spoken (e.g., posts comprising content in *Trentino* as spoken in the province of Trento – whose linguistic features exhibit traits of continuity between Lombard and Venetian [12] – are classified as Lombardy and Veneto, respectively).

Moreover, Friuli-Venezia Giulia and Apulia exhibit low scores on average across runs despite being represented in all data splits. The reason behind this has to be researched in linguistics rather than computation. Besides Friulian, Slovene and German varieties, in Friuli-Venezia Giulia varieties of Venetian are also spoken (e.g., the *Triestino* variety) [12], and thus posts comprising the latter are easily misclassified with the region in which Venetian is predominantly used (i.e., Veneto). On the other hand, *Salentino* varieties as spoken in the southern part of Apulia are part of the extreme southern varieties group [12], which also includes Sicilian, and thus make a large fraction of posts from Apulia to be misclassified as Sicily [8]. Besides the limitations of subtask A, this highlights that NLP should eventually go beyond "raw modeling" and start considering again linguistics as its foundation.

Subtask B: Fine-grained geolocation Test set results for all submitted runs in subtask B are reported in Table 3.

All teams except *SCG* outperformed both the baselines. The *baρtti* team obtained the best results with two out of three submissions (i.e., run 3 and 1). Their best run relies on multi-task learning on subtask A and B data, and uses geography-informed postprocessing to ensure that predictions fall inside the country borders. *DANTE*'s runs adopted similar methods to those employed in subtask A with separate layers for regression, ranking third with a model ensemble (run 3). *Salogni*'s run is based on UmBERTo fine-tuning, whereas the best run by *extremITA* is based on IT5 trained to generate region labels.

By looking at predictions by models that outperformed both baselines, we observe that, on average, errors range from 0.89 km to 668.11 km, with a median of 58.77 km. Errors are typically due to lexical items that are highly represented in other locations, e.g., posts with "*ghe* mel" (en: "of course", *Parmigiano* variety) fall in the Treviso area (Veneto) instead of the Parma area (Emilia-Romagna).

5.2.2. Special track

We present the results divided by subtask below.

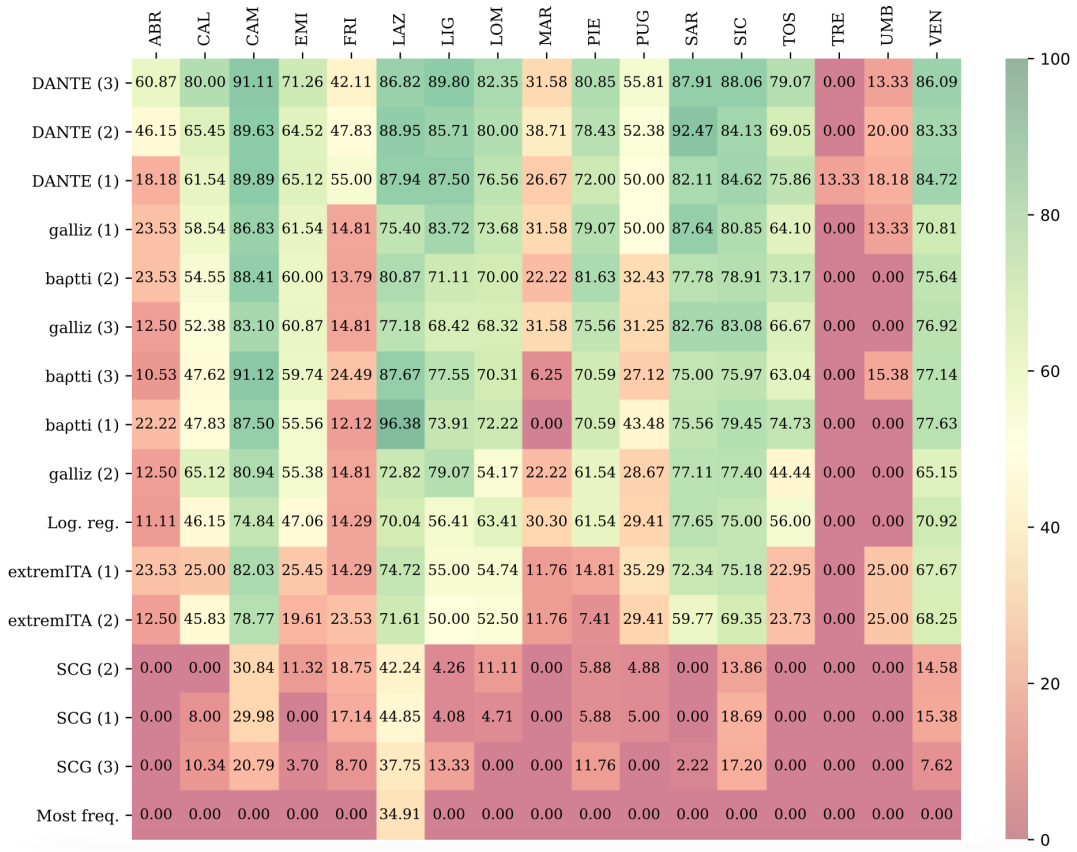


Figure 1: Results divided by region for the *standard track*, subtask A, in terms of F_1 score. Teams (with run numbers within parentheses) are on the rows, and test set regions (presented with their first three letters in Italian) are on the columns.

Subtask A: Coarse-grained geolocation Official results on the test set for the areas chosen by participant teams in subtask A (i.e., the *Tuscany-Lazio* area and the *Gallo-Italic* area) are summarized in Table 4.

As regards the *Tuscany-Lazio* area, the best run by the *galliz* team (run 3) achieved an improvement over the logistic regression baseline of 11.67 points in macro F_1 score. They employed a similar solution as the one for the *standard track*, additionally leveraging lexicons relevant to the linguistic area under consideration (i.e., lemmas from the *Vocabolario del Fiorentino Contemporaneo*[10] and a word list for the Romanesco dialect)[11] giving more weight to the BERT-based model. This confirms the usefulness of using region-specific linguistic materials in the task. For the *Gallo-Italic* area, all runs by the *SCG* team are between the two baselines we provided, but we are unfortunately unable to provide insights on their results.

[10]"Vocabolario del Fiorentino Contemporaneo" website: https://www.vocabolariofiorentino.it

[11]Romanesco word list from "The Roman Post" website: https://www.theromanpost.com/2016/06/dizionario-dialetto-romanesco

Table 4
Results on the test set for the *special track*, subtask A, divided by area. Baselines are italicized and highlighted in *yellow*.

	Team	Run	P	R	F_1
	Tuscany-Lazio area				
1	galliz	(3)	81.25	83.32	82.20
2	galliz	(1)	72.43	80.42	73.40
3	galliz	(2)	72.43	80.42	73.40
	Log. reg.		*91.79*	*66.67*	*70.53*
	Most freq.		*38.62*	*50.00*	*43.58*
	Gallo-Italic area				
	Log. reg.		*76.24*	*62.49*	*66.32*
1	SCG	(1)	30.86	31.36	29.14
2	SCG	(2)	29.42	29.03	26.66
3	SCG	(3)	25.30	26.25	22.78
	Most freq.		*10.06*	*25.00*	*14.35*

Table 5
Results on the test set for the *special track*, subtask B, divided by area. Baselines are italicized and highlighted in *yellow*.

	Team	Run	Avg dist (km)
	Gallo-Italic area		
	Centroid		*102.01*
	k-NN		*102.16*
1	SCG	(1)	102.41
2	SCG	(2)	104.59
3	SCG	(3)	111.79

Subtask B: Fine-grained geolocation In Table 5, we report the results for the area chosen by participants in subtask B. As for subtask A, we however do not have enough information to discuss further the *SCG*'s results.

6. Analysis and discussion

In this section, we analyze the approaches adopted by teams along several dimensions, providing a discussion and the insights derived from the shared task.

Models Apart from *SCG*, all participant teams used transformer-based language models for their runs. *Salogni* adopted an Italian RoBERTa-based model. *DANTE* and *baρtti* used versions of BERT pre-trained on Italian data, with the former using a much larger pre-training corpus than the latter, which might have impacted on the *DANTE* runs ranking first in subtask A. In contrast, *galliz* employed an English pre-trained BERT model, which still outperformed the logistic regression baseline in subtask A for both the tracks. This might indicate that subword tokenization in these models is suboptimal for the language varieties in DiatopIt, which naturally exhibits many non-Italian tokens with varied written forms, resulting in potentially small differences between Italian and English pre-trained models. Lastly, *extremITA* used a T5-based model pre-trained on Italian data and a LLaMA-based instruction-tuned model. Their results showed that recent large language models fine-tuned on disparate tasks are still far from tackling tasks such as GeoLingIt.

Multi-task learning Both *baρtti* and *DANTE* used multi-task learning in their submissions. While *baρtti* employed it during fine-tuning to exploit subtask A information to tackle subtask B and vice versa, *DANTE* used multi-task learning during a further stage of pre-training of a BERT-based model pre-trained on Italian data, which was then used to separately fine-tuning it on subtask A and B. Their pre-training setup consists of four tasks, including region-informed objectives, such as the prediction of the provenance region of posts and tokens. The approach followed by *DANTE* appears to lead to better performance in subtask A, whereas jointly training on both subtasks as done by *baρtti* seems to help in modeling fine-grained geolocation. Future work may shed light on how those approaches can help each other.

External resources Some participants used external resources to integrate the available data for the task. Three teams (i.e., *DANTE*, *galliz*, and *baρtti*) used data from a website containing a series of stories, poems, idioms, recipes, and articles in different language varieties that are spoken across Italy (i.e., Dialettando). In addition to this, *DANTE* also leveraged Wikipedia articles written in some of the language varieties that are present in our data. Both *DANTE* and *baρtti* used additional data from the Italian Wikipedia. For the *special track*, *galliz* also used lemmas from both a vocabulary of contemporary Florentine and a webpage for the Romanesco dialect (cf. Section 5.2.2). While *galliz* and *baρtti* used external data to create vocabularies, *DANTE* used it for pre-training their models. All of the teams who used external resources outperformed both baselines in both tasks, signaling that the use of external resources may indeed be pivotal in tackling the GeoLingIt task.

Data augmentation *baρtti* and *galliz* employed data augmentation techniques in order to artificially increase the amount of training data. *galliz* used external data to fine-tune an Italian word embeddings model, and then exploited it to swap randomly selected tokens with other semantically close ones. The *baρtti* team, on the other hand, constructed a vocabulary using external resources and then used it to randomly substitute tokens with other tokens from the vocabulary. Both teams outperformed our baselines, showing that the augmentation and diversification of training data can be useful for the task.

7. Conclusions

This paper provided an overview of GeoLingIt, the first shared task focused on geolocation of linguistic variation in Italy. The task attracted wide interest from the community, registering 37 expressions of interest and 35 official runs. After presenting participants' results and the adopted approaches, we outlined the main insights from the shared task. Besides natural language processing, we hope that GeoLingIt sensitized the community on the linguistic diversity of the country.

References

[1] M. Maiden, M. Parry, The dialects of Italy, Romance Linguistics, Routledge, 1997.

[2] A. Ramponi, NLP for language varieties of Italy: Challenges and the path forward, arXiv preprint arXiv:2209.09757 (2022). URL: https://arxiv.org/abs/2209.09757.

[3] F. Avolio, Lingue e dialetti d'Italia, Le Bussole, Carocci, Roma, Italy, 2009.

[4] J. Eisenstein, What to do about bad language on the internet, in: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Atlanta, Georgia, 2013, pp. 359–369.

[5] R. van der Goot, A. Ramponi, A. Zubiaga, B. Plank, B. Muller, I. San Vicente Roncal, N. Ljubešić, Ö. Çetinoğlu, R. Mahendra, T. Çolakoğlu, T. Baldwin, T. Caselli, W. Sidorenko, MultiLexNorm: A shared task on multilingual lexical normalization, in: Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021), Association for Computational Linguistics, Online, 2021, pp. 493–509. URL: https://aclanthology.org/2021.wnut-1.55. doi:10.18653/v1/2021.wnut-1.55.

[6] G. Berruto, Quale dialetto per l'Italia del duemila? Aspetti dell'italianizzazione e risorgenze dialettali in Piemonte (e altrove), in: Lingua e dialetto nell'Italia del Duemila, Congedo, 2006, pp. 101–127.

[7] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[8] A. Ramponi, C. Casula, DiatopIt: A corpus of social media posts for the study of diatopic language variation in Italy, in: Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023), Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 187–199. URL: https://aclanthology.org/2023.vardial-1.19.

[9] B. Han, A. Rahimi, L. Derczynski, T. Baldwin, Twitter geolocation prediction shared task of the 2016 workshop on noisy user-generated text, in: Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT), The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 213–217.

[10] M. Gaman, D. Hovy, R. T. Ionescu, H. Jauhiainen, T. Jauhiainen, K. Lindén, N. Ljubešić, N. Partanen, C. Purschke, Y. Scherrer, M. Zampieri, A report on the VarDial evaluation campaign 2020, in: Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects, International Committee on Computational Linguistics (ICCL), Barcelona, Spain (Online), 2020, pp. 1–14.

[11] B. R. Chakravarthi, G. Mihaela, R. T. Ionescu, H. Jauhiainen, T. Jauhiainen, K. Lindén, N. Ljubešić, N. Partanen, R. Priyadharshini, C. Purschke, E. Rajagopal, Y. Scherrer, M. Zampieri, Findings of the VarDial evaluation campaign 2021, in: Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects, Association for Computational Linguistics, Kiyv, Ukraine, 2021, pp. 1–11.

[12] G. B. Pellegrini, Carta dei dialetti d'Italia, Profilo dei Dialetti Italiani, Pacini, Pisa, Italy, 1977.

[13] A. Ramponi, S. Tonelli, Features or spurious artifacts? data-centric baselines for fair and robust hate speech detection, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 3027–3040. URL: https://aclanthology.org/2022.naacl-main.221. doi:10.18653/v1/2022.naacl-main.221.

[14] A. Koudounas, F. Giobergia, I. Benedetto, S. Monaco, L. Cagliero, D. Apiletti, E. Baralis, baρtti at GeoLingIt: Beyond boundaries, enhancing geolocation prediction and dialect classification on social media in Italy, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[15] G. Gallipoli, M. La Quatra, D. Rege Cambrin, S. Greco, L. Cagliero, DANTE at GeoLingIt: Dialect-aware multi-granularity pre-training for locating tweets within Italy, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[16] C. D. Hromei, D. Croce, V. Basile, R. Basili, ExtremITA at EVALITA 2023: Multi-task sustainable scaling to large language models at its extreme, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[17] T. Labruna, S. Gallo, Galliz at GeoLingIt: Enhancing BERT with vocabulary knowledge for predicting the region of language varieties of Italy, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[18] I. Salogni, Salogni at GeoLingIt: Geolocalization by fine-tuning BERT, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

DANTE at GeoLingIt: Dialect-Aware Multi-Granularity Pre-training for Locating Tweets within Italy

Giuseppe Gallipoli[1,†], Moreno La Quatra[2,†], Daniele Rege Cambrin[1,†], Salvatore Greco[1,†] and Luca Cagliero[1]

[1]Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Turin TO
[2]Kore University of Enna, Piazza dell'Università, 94100 Enna EN

Abstract

This paper presents an NLP research system designed to geolocate tweets within Italy, a country renowned for its diverse linguistic landscape. Our methodology consists of a two-step process involving pre-training and fine-tuning phases. In the pre-training step, we take a semi-supervised approach and introduce two additional tasks. The primary objective of these tasks is to provide the language model with comprehensive knowledge of language varieties, focusing on both the sentence and token levels. Subsequently, during the fine-tuning phase, the model is adapted explicitly for two subtasks: coarse- and fine-grained variety geolocation. To evaluate the effectiveness of our methodology, we participate in the GeoLingIt 2023 shared task and assess our model's performance using standard metrics. Ablation studies demonstrate the crucial role of the pre-training step in enhancing the model's performance on both tasks.

Keywords

Linguistic Varieties, Region Localization, Text Classification and Regression, Italian NLP

1. Introduction

Italy is widely recognized for its linguistic diversity, with 20 distinct regions, each characterized by various unique and shared dialects [1]. These dialects exhibit further variations within each region, often associated with specific cities or provinces, and sometimes extend beyond regional boundaries. The intricate nature of Italy's linguistic landscape poses a significant challenge in accurately identifying the origin of a given text within the country.

This research is conducted in the context of the GeoLingIt shared task [2] at EVALITA 2023 [3]. It focuses on the geolocation of social media data, specifically Twitter posts. The task comprises two subtasks: *Coarse-grained variety geolocation* (Subtask A), whose aim is to determine the region from which a tweet originates within the 20 Italian regions, and *Fine-grained variety geolocation* (Subtask B), which focuses on predicting the latitude and longitude coordinates corresponding to the origin of a tweet within Italy. Linguistic variations within and across regions make it difficult to accurately associate a piece of text with its specific geographic origin. The challenge becomes even more significant due to the similarities each language variety may share with other languages, even outside Italy.

This paper presents the DANTE (Dialect ANalysis TEam)[1] submission for the GeoLingIt 2023 shared task, characterized by a two-step methodology involving pre-training and fine-tuning phases. By leveraging Italian or multilingual models, we propose a semi-supervised pre-training approach that combines standard and novel pre-training tasks to capture regional dialect information at multiple levels of granularity (i.e., sentence and token levels). Following the pre-training phase, the model undergoes a standard fine-tuning process tailored to the two subtasks proposed by the shared task. Through extensive experiments, we demonstrate the effectiveness of our methodology.

2. Background

Text classification is a fundamental task in NLP whose objective is to assign one (or more) predefined classes to a piece of text. It has many applications ranging from sentiment analysis to topic classification. In this work, we apply it to the prediction of the geographic region associated with the linguistic variety expressed in a tweet.

[1]The name "DANTE" is inspired by the Italian poet Dante Alighieri, widely regarded as one of the founding fathers of the Italian language.

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT

† Corresponding authors. These authors contributed equally.
giuseppe.gallipoli@polito.it (G. Gallipoli); moreno.laquatra@unikore.it (M. L. Quatra); daniele.regecambrin@polito.it (D. R. Cambrin); salvatore_greco@polito.it (S. Greco); luca.cagliero@polito.it (L. Cagliero)
https://mlaquatra.me (M. L. Quatra); https://darthreca.github.io/ (D. R. Cambrin); https://grecosalvatore.github.io/ (S. Greco)
0009-0003-1744-6674 (G. Gallipoli); 0000-0001-8838-064X (M. L. Quatra); 0000-0002-5067-2118 (D. R. Cambrin); 0000-0001-7239-9602 (S. Greco); 0000-0002-7185-5247 (L. Cagliero)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper14.pdf

The introduction of the Transformer [4] architecture for machine translation has represented a significant breakthrough in NLP, achieving superior performance also in other tasks, including text classification. Transformer-based classification models implement an encoder-only architecture whose objective is to extract a continuous representation from the input text. To do this, models are generally pre-trained on large corpora of unlabeled text using specific pre-training objectives.

The pre-training stage allows the model to learn language representations that enable it to capture the structure and semantics of the text more effectively. Our work follows the same approach by further pre-training several Transformer-based models as discussed in Section 3. After pre-training, the model is fine-tuned on labeled data tailored to the desired task. Specifically, the architecture is enriched by additional classification layers (i.e., classification head) trained in a supervised fashion to output the final probability for each class. Similarly, by introducing one or multiple linear layers, (multi-)regression tasks can also be performed.

Some of the most widely adopted Transformer-based classification models include: BERT and its multilingual version mBERT [5], DistilBERT [6], which is a distilled version of BERT, RoBERTa [7], and its multilingual version XLM [8], which are two variations of BERT including dynamic masking.

Computational linguistics research in Italian faces challenges due to the scarcity of large-scale datasets specifically designed for the language, as highlighted recently [9]. Also, the computational effort required to pre-train language models has resulted in only a few available architectures in Italian. Specifically, some of them are BERT-Italian and ELECTRA-Italian [10]. Furthermore, although they are not encoder-only architectures, the following are some of the other models available in Italian: GePpeTto [11], which is based on GPT-2 [12], IT5 [13], which is the Italian version of T5 [14], and the recently released BART-IT [15], which is the Italian version of BART [16].

3. Description of the system

The DANTE methodology for the GeoLingIt shared task aims to both identify the region of origin and predict the geographic coordinates of tweets within Italy.

3.1. Pre-training

The initial phase of our methodology involves pre-training the model to improve its ability to analyze different linguistic varieties. We initialize a Transformer-based encoder model using Italian or multilingual pre-trained weights and further pre-train it using both standard and novel pre-training tasks.

Masked Language Modeling (MLM) & Next Sentence Prediction (NSP). The MLM and NSP tasks are standard pre-training tasks used to train Transformer-based models. Both tasks contribute to language processing by helping the model learn the contextual information of words and their relationships.

Region Classification (RC). By leveraging region-specific data, we integrate into pre-training the supervised task of predicting the geographic region associated with the linguistic variety expressed in a given sentence.

Token-level Region Classification (TRC). We also include an additional (supervised) token classification task. It aims at predicting the geographic region associated with each token in a given sentence. To create training examples, we randomly combine multiple sentences belonging to text snippets labeled with different regions. This task aims at enabling the model to capture regional linguistic information with higher granularity.

Using a multi-task learning approach, the model is trained on multiple tasks simultaneously, allowing it to learn a shared representation useful for all tasks. We define a separate linear layer for each task (i.e., task-specific head) that operates on the shared representation and is trained using the corresponding labeled data. We experimented with two different multi-task learning setups: (1) *task-specific training (TST)*, where the model is trained on a single task at a time, with each batch randomly selecting one task from the set of all available tasks, and (2) *joint training (JT)*, where the model is trained on all tasks simultaneously, and the loss is computed as the average of the losses of all tasks. These two multi-task learning setups were inspired by recent findings in the literature [17].

3.2. Subtask A: Coarse-grained variety geolocation

The *Subtask A* within GeoLingIt 2023 shared task involves identifying the region of origin of a given tweet within Italy. It can be formulated as a classification task, where the model is trained to classify each tweet into its corresponding geographic region (i.e., one of the 20 Italian regions). To this end, we follow a standard fine-tuning approach, where the pre-trained model is adapted for the downstream task using the labeled training data. The representation of a special `[CLS]` token is used as the input to a linear layer trained to predict the region of origin of the tweet. The model is trained to minimize

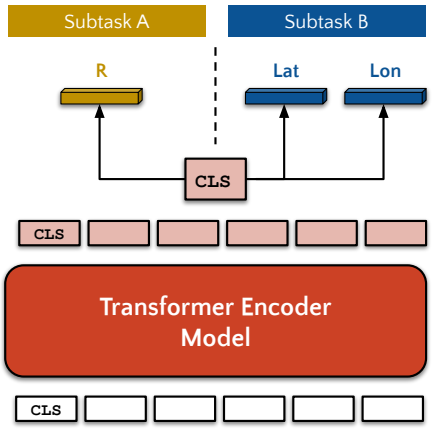


Figure 1: Fine-tuning architecture. It includes two branches: one for *Subtask A* which predicts the region class (represented as "R"), and another for *Subtask B* which predicts the latitude (represented as "Lat") and longitude (represented as "Lon").

the cross-entropy loss between the predicted and the ground-truth labels. For a visual representation of the fine-tuning process, please refer to the top-left part of Figure 1.

3.3. Subtask B: Fine-grained variety geolocation

The *Subtask B* aims to localize a given tweet's origin within Italy by predicting its latitude and longitude coordinates. The task can be formulated as a regression problem, where the model is trained to jointly predict two values (i.e., the latitude and longitude coordinates of the tweet's origin) using two separate linear layers. Similar to the previous case, the tweet representation is obtained by feeding the `[CLS]` token representation to both linear layers. The fine-tuning architecture for this subtask is illustrated in the upper right corner of Figure 1, showcasing the linear layers positioned on top of the `[CLS]` token.

The overall loss function for the regression task is defined as:

$$\mathcal{L}_{localization} = \frac{1}{2}\left(\mathcal{L}_{lat} + \mathcal{L}_{lon}\right)$$

Where $\mathcal{L}_{lat}$ and $\mathcal{L}_{lon}$ represent the mean squared error (MSE) loss for the latitude and longitude predictions, respectively.

It is worth noting that the model is separately fine-tuned for each task (i.e., coarse- and fine-grained variety geolocation). Thus, we do not use multi-task learning at this stage. Jointly optimizing the two tasks during fine-tuning could help the model to ensure consistency between the two tasks and improve the model's performance. This is one of the possible future directions we plan to explore.

4. Dataset

4.1. Pre-training Dataset

To the best of our knowledge, there are no existing large-scale data collections specifically focusing on Italian language varieties. Therefore, we exploited web scraping to construct our pre-training dataset. From a web search, we identified the following two sources: (1) Dialettando[2]: a website that contains several proverbs, sayings, poems, rhymes, and stories from different regions; (2) Wikipedia: which comprises specific versions for some regional languages (e.g., nap[3] for Neapolitan). They were both accessed in January 2023. For the data collected from Wikipedia, we associated each language-specific Wikipedia portal with the region primarily representing the respective language. For example, data collected from the nap Wikipedia portal would be associated with the Campania region, which predominantly represents the Neapolitan language. After the data collection, we ended up with a corpus of 273,011 documents containing linguistic varieties of Italian from different regions. Out of these, 12,692 documents were collected from Dialettando, while the majority, 260,319, were obtained from Wikipedia. From Dialettando, we also collected the 12,692 Italian translations of the same documents in the corpus. This was done because the DiatopIt corpus utilized in the task contains instances that encompass regional Italian variations. Therefore, the final pre-training dataset is composed of 285,703 documents. Notice that a document can be a Wikipedia article or any text from Dialettando (e.g., proverb, saying, or story) without any difference, even if they can have different lengths. Indeed, we found that the mean number of tokens is 48 for Dialettando and 147 for Wikipedia[4]. However, both sources of texts can be helpful during the pre-training phase.

Figure 2 shows the distribution of the collected documents for each Italian region. Table 1 details the number of documents for each region and data source. As can be noticed, regions of the north of Italy, such as Piemonte, Lombardia, and Veneto, are predominant in the dataset, with approximately 60k texts (corresponding to approximately 20% of the entire collection) each. They are followed by some regions of the south, such as Sicilia, Campania, and Puglia, with around 25k, 14k, and 11k texts, respectively. Finally, regions such as Valle D'Aosta, Toscana, Umbria, Marche, Lazio, Molise, Abruzzo, and

[2] https://www.dialettando.com
[3] https://nap.wikipedia.org/
[4] Computed with the `bert-base-multilingual-cased` tokenizer of the HuggingFace library https://huggingface.co/bert-base-multilingual-cased.

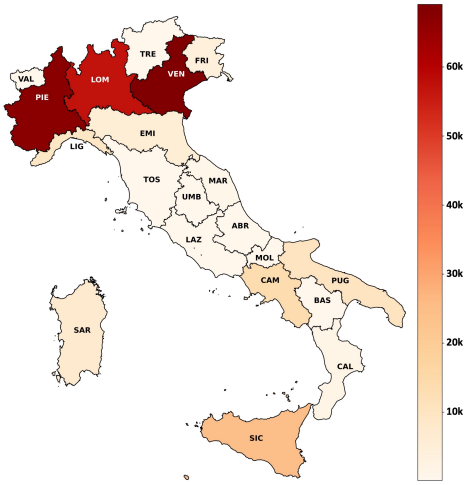


Figure 2: Distribution of the collected pre-training texts per region.

Region	Abbr.	# Documents Dial.	Wiki
Abruzzo	ABR	325	–
Basilicata	BAS	260	–
Calabria	CAL	1,258	–
Campania	CAM	1,340	12,709
Emilia Romagna	EMI	492	5,194
Friuli Venezia Giulia	FRI	157	3,833
Lazio	LAZ	356	–
Liguria	LIG	1,667	7,838
Lombardia	LOM	1,332	55,743
Marche	MAR	690	–
Molise	MOL	153	–
Piemonte	PIE	854	66,407
Puglia	PUG	1,504	9,063
Sardegna	SAR	296	7,262
Sicilia	SIC	843	23,732
Toscana	TOS	242	–
Trentino Alto Adige	TRE	127	–
Umbria	UMB	161	–
Valle D'Aosta	VAL	102	–
Veneto	VEN	533	68,538
Total		*12,692*	*260,319*

Table 1
Number of collected documents for each Italian region. For each Italian region is reported: i) the abbreviation code; ii) the number of documents collected from Dialettando and Wikipedia. The total number of documents containing Italian language varieties from all regions is is 273,011.

Basilicata are lowly represented, with less than 1k texts each.

Notably, the data utilized for pre-training is sourced from the Dialettando website and Wikipedia, adhering to the guidelines of the GeoLingIt shared task by excluding social network data to avoid overlap with the fine-tuning data.

4.2. GeoLingIt Dataset

The GeoLingIt 2023 dataset for the *Subtasks A* and *B* is DiatopIt [18], a corpus of diatopic variations of language in Italy. It is composed of geotagged social media posts from Twitter. Each tweet also comprises the associated latitude, longitude, and the Italian region of origin. The dataset contains 13,669 examples for training, 552 for validation, and 818 for testing. This dataset is exploited in the fine-tuning phase to specialize the models to coarse- and fine-grained variety geolocation. The authors of the competition have already anonymized data. Specifically, user mentions, email addresses, and URLs have been replaced with specific placeholders for privacy reasons. However, the content of tweets is unfiltered and can exhibit non-standard language use (e.g., insults, bad words).

5. Results

Baseline models We compare our models with baseline models pre-trained on Italian or multilingual data, which undergo the same fine-tuning process described in Sections 3.2 and 3.3. The following baseline models are considered: (1) multilingual BERT model (mBERT)[5], (2) BERT model pre-trained on 13GB of Italian text (BERT-IT)[6], (3) BERT model pre-trained on 81GB of Italian text (BERT-IT-XXL)[7], (4) XLM-RoBERTa (XLM-R)[8], and (5) multilingual DistilBERT (dBERT)[9]. All models are used in their cased versions. By comparing the results of these baseline models with our approach, we can assess the benefits of the proposed pre-training phase.

Experimental settings The pre-training phase lasts five epochs and utilizes the Adam optimizer with a linear learning rate scheduler. The scheduler includes a warmup period (10% of the total training steps) followed by a linear decay of the learning rate until the end of training. The fine-tuning phase lasts for ten epochs and utilizes the same settings for the optimizer and scheduler as the pre-training phase.

[5] https://huggingface.co/bert-base-multilingual-cased
[6] https://huggingface.co/dbmdz/bert-base-italian-cased
[7] https://huggingface.co/dbmdz/bert-base-italian-xxl-cased
[8] https://huggingface.co/xlm-roberta-base
[9] https://huggingface.co/distilbert-base-multilingual-cased

5.1. Coarse-grained variety geolocation

We report the single models and an ensemble, which includes all models evaluated on the development set. The ensemble prediction is obtained through majority voting on the individual models' predictions. In case of a tie, a random selection is employed. The organizers provide Logistic Regression (LogReg) and Most Frequent Baseline (MFB) as baselines. According to the competition rules, we consider macro F1-score, precision, and recall as evaluation metrics.

In Table 2, we reported development and test sets results. There is a noticeable performance improvement when comparing deep learning techniques to classical machine learning methods. One common aspect shared by both approaches is the observed degradation in performance on the test set compared to the development set. This pattern can be attributed to the fact that the test set contains additional regions that are not present in the development set. Joint-Training (JT) consistently yields the best results in terms of F1-score, achieving significant improvements ranging from +2% to +7% compared to the absence of pre-training. This boost primarily manifests as enhanced precision.

Following GeoLingIt guidelines, we can only submit three models for test set evaluation. We have selected the top-3 models based on their performance on the development set: Jointly-Trained BERT-IT-XXL, Task-Specific-Trained BERT-IT-XXL, and the models' Ensemble. We show in Table 2b the performance of these models on the test set. The results show that the Ensemble method achieved the highest performance, followed by the Task-Specific-Trained (TST) BERT-IT-XXL model and the Jointly-Trained (JT) BERT-IT-XXL model. Surprisingly, the TST pre-training approach outperformed the others, exhibiting a significant +2% improvement in F1-score compared to the corresponding model pre-trained using JT.

Model	PT	Precision	Recall	F1-score
LogReg	✗	0.7686	0.5389	0.5872
MFB	✗	0.0160	0.0769	0.0265
BERT-IT-XXL	✗	0.7498	0.7914	0.7260
	JT	0.8401	0.7638	0.7923
	TST	0.8101	0.7374	0.7621
BERT-IT	✗	0.7276	0.7795	0.7005
	JT	0.7757	**0.8161**	0.7496
	TST	0.7591	0.8075	0.7310
dBERT	✗	0.7032	0.7555	0.6747
	JT	0.7196	0.7651	0.6962
	TST	0.7202	0.7830	0.6941
mBERT	✗	0.7096	0.7476	0.6874
	JT	0.7358	0.7872	0.7222
	TST	0.7408	0.7999	0.7170
XLM-R	✗	0.6856	0.7257	0.6723
	JT	0.7642	0.8112	0.7438
	TST	0.7325	0.7709	0.7136
Ensemble	E	**0.8609**	0.7878	**0.8121**

(a) Development Set

Model	PT	Precision	Recall	F1-score
LogReg	✗	0.6219	0.4243	0.4611
MFB	✗	0.0447	0.2115	0.0738
BERT-IT-XXL	JT	0.6518	0.6009	0.6172
BERT-IT-XXL	TST	0.6698	0.6265	0.6393
Ensemble	E	**0.7946**	**0.6375**	**0.6630**

(b) Test Set

Table 2
Subtask A results. The *PT* column indicates no-pre-training with ✗, Task-Specific Training with *TST*, Joint Training with *JT*, and models ensemble with *E*.

5.2. Fine-grained variety geolocation

Task organizers provide K-nearest-neighbor (KNN) and centroid baseline (CB) models as baselines. Following shared task guidelines, the models' performance is assessed using haversine distance. In this case, the lower, the better. We report the results of single models and an ensemble of the top-2 evaluated models. The ensemble prediction is obtained using the mean point between the two individual models' predictions.

In Table 3, we reported results for the evaluated model on the development and test sets. Similar to Subtask A, deep learning models outperform classical approaches. Notably, the test set's performance in this task shows higher scores than the development set.

The results on the development set confirm the effectiveness of the pre-training with one exception: Task-Specific-Training on BERT-IT. The differences span from -19 km to -161 km with respect to the models without a specific pre-training. In most cases, Joint-Training consistently yields the best results, except for XLM-R.

We submit the top-3 most promising solutions according to the development set results for evaluation on the test set: Jointly-Trained dBERT, Task-Specific-Trained dBERT, and Ensemble. We report their performance on the test set in Table 3b. The Ensemble model achieved the best performance, with the TST pre-training demonstrating a 1.5 km average distance improvement compared to the JT counterpart.

Model	PT	Avg Dist (km)
KNN	✗	281.03
CB	✗	301.65
BERT-IT-XXL	✗	161.88
	JT	128.18
	TST	136.00
BERT-IT	✗	187.47
	JT	153.26
	TST	215.35
dBERT	✗	143.61
	JT	118.57
	TST	124.15
mBERT	✗	192.20
	JT	135.04
	TST	164.99
XLM-R	✗	316.90
	JT	156.86
	TST	155.83
Ensemble	E	**117.64**

(a) Development Set

Model	PT	Avg dist (km)
KNN	✗	263.35
CB	✗	281.04
dBERT	JT	114.00
dBERT	TST	112.58
Ensemble	E	**110.35**

(b) Test Set

Table 3
Subtask B results. The *PT* column indicates no-pre-training with ✗, Task-Specific Training with *TST*, Joint Training with *JT*, and models ensemble with *E*.

6. Discussion

Our analysis assessed the effectiveness of widely used deep language models in the context of both coarse- and fine-grained variety geolocation tasks. We also offer an interactive demo[10] to showcase our best-performing models and release both the code and pre-trained models[11]. It is worth noting that social media data composing the fine-tuning dataset may contain profanities, slurs, hateful content, and stereotypes. Although pre-training data is collected using controlled sources, a similar statement may apply to them. A community partially manages both the Dialettando website and Wikipedia portals. Therefore their content may not be carefully curated. As a result, the models may exhibit label correlations based on the presence of such offensive language, potentially influencing their region identification capabilities. The proposed methodologies are not intended to offend anyone; since they may be inaccurate in some cases, it is possible to get improprieties.

[10] https://huggingface.co/spaces/DGMS/DANTE-GeoLingIT2023
[11] https://github.com/MorenoLaQuatra/DANTE-GeoLingIT2023

7. Conclusion and Future Works

This paper presents an effective solution for modeling language varieties within Italy, achieving excellent results and ranking 1st and 2nd among other teams for Subtask A and Subtask B, respectively. However, there are still promising avenues for future research. Utilizing multi-task learning during the fine-tuning phase can improve consistency and performance by training on multiple related tasks using the same backbone model. Regarding model architecture, we aim to investigate the development of a specific model focused on identifying portions of the text belonging to specific language varieties. This model will be designed to identify the distinctive linguistic features within tweets accurately. By successfully identifying these features, the model would have the potential to concentrate on the relevant parts of the text, which may lead to improved localization capabilities. Finally, preliminary experiments show that incorporating curriculum learning techniques during pre-training can optimize the learning process and enhance the overall model's performance.

Acknowledgments

This study was partially carried out within the MICS (Made in Italy – Circular and Sustainable) Extended Partnership and received funding from Next-GenerationEU (Italian PNRR – M4 C2, Invest 1.3 – D.D. 1551.11-10-2022, PE00000004). This study was also partially carried out within the FAIR (Future Artificial Intelligence Research) and received funding from Next-GenerationEU (Italian PNRR – M4 C2, Invest 1.3 – D.D. 1555.11-10-2022, PE00000013). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

References

[1] A. Ramponi, Nlp for language varieties of italy: Challenges and the path forward, arXiv preprint arXiv:2209.09757 (2022).
[2] A. Ramponi, C. Casula, GeoLingIt at EVALITA 2023: Overview of the geolocation of linguistic variation in Italy task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and

Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[3] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[5] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: 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), 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[6] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, ArXiv abs/1910.01108 (2019).

[7] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692.

[8] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440–8451. URL: https://aclanthology.org/2020.acl-main.747. doi:10.18653/v1/2020.acl-main.747.

[9] A. Koudounas, M. La Quatra, L. Vaiani, L. Colomba, G. Attanasio, E. Pastor, L. Cagliero, E. Baralis, Italic: An italian intent classification dataset, arXiv preprint arXiv:2306.08502 (2023).

[10] S. Schweter, Italian bert and electra models, 2020. URL: https://doi.org/10.5281/zenodo.4263142. doi:10.5281/zenodo.4263142.

[11] L. D. Mattei, M. Cafagna, F. Dell'Orletta, M. Nissim, M. Guerini, Geppetto carves italian into a language model, 2020. arXiv:2004.14253.

[12] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, 2019.

[13] G. Sarti, M. Nissim, IT5: Large-scale text-to-text pretraining for italian language understanding and generation, ArXiv preprint 2203.03759 (2022). URL: https://arxiv.org/abs/2203.03759.

[14] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (2020).

[15] M. La Quatra, L. Cagliero, BART-IT: An Efficient Sequence-to-Sequence Model for Italian Text Summarization, Future Internet 15 (2022) 15.

[16] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 7871–7880. URL: https://aclanthology.org/2020.acl-main.703. doi:10.18653/v1/2020.acl-main.703.

[17] C. Fifty, E. Amid, Z. Zhao, T. Yu, R. Anil, C. Finn, Efficiently identifying task groupings for multi-task learning, Advances in Neural Information Processing Systems 34 (2021) 27503–27516.

[18] A. Ramponi, C. Casula, DiatopIt: A corpus of social media posts for the study of diatopic language variation in Italy, in: Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023), 2023, pp. 187–199. URL: https://aclanthology.org/2023.vardial-1.19.

Galliz at GeoLingIt: Enhancing BERT with Vocabulary Knowledge for Predicting the Region of Language Varieties of Italy

Tiziano Labruna[1,2], Simone Gallo[3,4]

[1]Fondazione Bruno Kessler, Trento, Italy
[2]Free University of Bozen-Bolzano, Italy
[3]HIIS Laboratory, CNR - ISTI, Pisa, Italy
[4]University of Pisa, Italy

Abstract
The linguistic diversity of the Italian peninsula and its islands, characterized by several language varieties, represents a linguistic condition and a cultural treasure unique in Europe. However, the oral nature of these varieties poses a challenge to their preservation in the written form. While significant research efforts have been dedicated to standard Italian language processing, less attention has been given to the language varieties of Italy and the development of supporting resources. This paper aims to study the peculiarities of language varieties of Italy and identify the region of origin of tweets written in *non-[Standard Italian]* varieties. To achieve this goal, we utilized two main techniques: fine-tuning a language model (BERT) and implementing an algorithm that utilizes dictionaries of regional varieties and word frequency. Our results show that integrating lexical analysis with BERT could be a promising approach for this particular task. We present an overview of the data, methodology, and evaluation results, then discuss the implications of our findings.

Keywords
Natural Language Processing, Language varieties, Tweets classification

1. Introduction and Motivations

The Italian peninsula and its islands present considerable linguistic variation among the different regions that compose it, as well as within the regions themselves. The presence of many different language varieties makes this linguistic situation special and unique in Europe [1], as well as a treasure of cultural diversity, interpretation, and expression of the reality to which they belong. However, these linguistic diversities are in danger of being lost, as most of them are passed on only orally, leaving less room for written usage [1]. Despite significant research efforts being devoted to processing techniques for standard Italian (e.g., [2, 3, 4]), less effort has been devoted to supporting language varieties, both from a technological point of view and in terms of curated resources [5]. In this paper, our main goal is to study varieties of Italy in order to develop effective methods for classifying the region of origin of Twitter posts (tweets) written from Italy. We address two different tasks of GeoLingIt 2023 [6] from EVALITA 2023 [7]: one for classifying non-standard Italian tweets according to their region of origin at the country level (*"standard track"*), and another for classifying tweets according to a subset of regions (*"special track"*, in this case, Lazio and Toscana). To tackle this problem, we rely on the combination of two different techniques: the first one is based on the fine-tuning of a language model (i.e., BERT), while the second is based on an algorithm that utilizes regional varieties dictionaries and the frequency of words present in the tweets. The classification results obtained from both techniques are normalized and combined to derive the final result. In the following sections, we provide an overview of the data and resources used for the tasks. We then describe the methodology applied, including data augmentation, prediction using the two different techniques, and global prediction. Next, we present the results obtained during the evaluation phase. Finally, we discuss the findings and draw some conclusions.

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
tlabruna@fbk.eu (T. Labruna); simone.gallo@isti.cnr.it (S. Gallo)
0000-0001-7713-7679 (T. Labruna); 0000-0002-5162-0475 (S. Gallo)
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

2. Data and Resources

The dataset used for the tasks was collected by retrieving geotagged tweets classified as *IT* by Twitter. The curators only kept posts that exhibit *non-standard* language, along with the region information that falls within the Italy territory [6].

2.1. Provided Data

The provided data consists of training and development splits for both the *standard track* and the *special track*

CEUR-WS.org/Vol-3473/paper15.pdf

tasks. The data is provided in a tab-separated format, with each column defining three properties:

- "id": an integer that uniquely identifies the tweet;
- "text": a string representing the text of the tweet in a *non-standard* language variety of Italy. This variety may be present as a single word or phrase (there are many cases of code-switching), or the entire tweet can be written in that variety. Any sensitive information has been replaced with placeholders by the curators (e.g., "@tagged_user" is replaced with "[USER]");
- "region": a string representing the tweets' region of provenance (e.g., Lazio, Sicilia, Toscana).

The training set contains 13,669 tweets, covering all the administrative regions of Italy. The development set consists of 552 tweets from 13 selected regions, namely Calabria, Campania, Friuli-Venezia Giulia, Emilia Romagna, Lazio, Liguria, Lombardia, Piemonte, Puglia, Sardegna, Sicilia, Toscana, and Veneto. The training set exhibits a strong imbalance, with highly represented regions like Lazio (5549 items) and Campania (2971 items), while regions such as Valle d'Aosta and Molise have only 14 and 35 items, respectively. The overall class distribution is shown in Fig. 1.

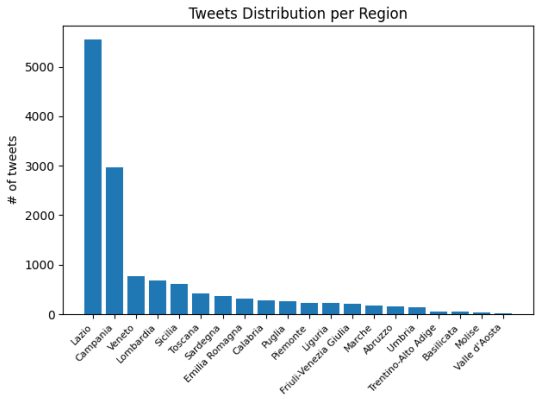


Figure 1: Distribution of tweets for each region in the training set.

2.2. Language Model

We employed BERT (Bidirectional Encoder Representations from Transformers) [8] as our language model, which was introduced by Google in 2018 and has gained significant prominence in various Natural Language Processing (NLP) tasks. Unlike traditional language models that utilize left-to-right or right-to-left approaches, BERT utilizes bidirectional pre-training, allowing all tokens in the input to contribute to the prediction process. Its effectiveness is evident from its state-of-the-art performance on multiple NLP benchmarks, including GLUE, SQuAD and RACE.

Although BERT has been in existence for five years, we believe that it remains highly suitable for our specific task. In light of its robust performance across diverse NLP applications, leveraging BERT finely aligns with the requirements of our work.

During the model selection, we also took into account Italian variants of the classic BERT model (e.g., "bert-base-italian"[1]), but the preliminary results reported, in some case, worst performance compared to the "standard" English BERT (i.e., "bert-base-uncased"[2])

2.3. Vocabularies

To build the vocabularies needed for our purpose, we utilized various online resources as well as the text from the provided tweets. Some of these vocabularies were used for the *standard track*, while other ones for the *special track*, as more precisely described in the rest of this Section. All vocabularies are publicly available[3].

Global vocabulary We obtained a "global" vocabulary, containing words from language varieties spoken in every Italian region, by performing web scraping on the dictionary available at "Dialettando.com"[4]. This resulted in a JSON file containing all the available words for each region. This vocabulary was used for the *standard track* only.

Unique words vocabulary This vocabulary was generated starting from the provided training set and considering the occurrences of the words for every tweet from each region. Specifically, it contains all the unique words present in the tweets, along with their corresponding frequencies, grouped by region. This vocabulary was used for the *standard track*, and a subset of this vocabulary with only the regions Toscana and Lazio was used for the *special track*.

Distinctive words vocabulary Similarly to the previous one, this vocabulary was generated from the provided training set. This time, we keep only the distinctive words from each region's tweets (i.e., only words that are exclusive to a specific region and do not appear in other regions are considered), along with the corresponding frequencies. This vocabulary was used for the *standard*

[1]https://huggingface.co/dbmdz/bert-base-italian-cased
[2]https://huggingface.co/bert-base-uncased
[3]https://github.com/simog-dev/Galliz_geolingit
[4]https://www.dialettando.com/dizionario/dizionario.lasso, *retrieved on May 9th, 2023*

track, and a subset of this vocabulary with only the regions Toscana and Lazio was used for the *special track*.

Toscana vocabulary We obtained this vocabulary by performing a web scraping of the terms present in the website of "Vocabolario del Fiorentino Contemporaneo"[5] and thus converting the content of the website into a JSON file. This vocabulary was used for the *special track* only.

Lazio vocabulary We obtained this vocabulary by performing a web scraping of the terms present in the website of "The Roman Post" website[6] and thus converting the content of the website into a JSON file. This vocabulary was used for the *special track* only.

3. Methodology

In order to predict the most likely region of origin of a given sentence, we decided to make use of both LLMs (fine-tuned on an augmented training set) and lexical information of regional varieties, taken from the vocabularies presented above. As shown in Fig. 2, we first consider the predictions of the two strategies individually, and then we merge both contributions for coming to one final prediction of the region enclosing the particular variety for the sentence.

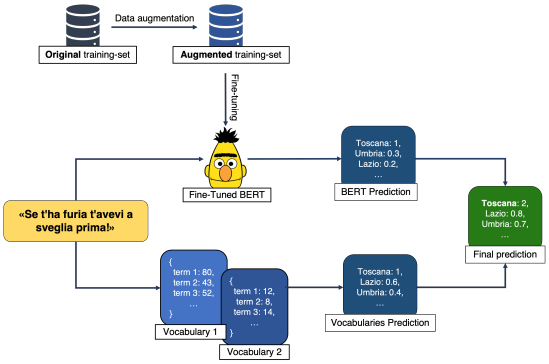


Figure 2: System Architecture

3.1. Data Augmentation

The first step was to increase the size of the training set, both for producing better training of the classifier, and for equalizing the distribution of the samples for each region, which was initially extremely unbalanced (cfr. § 2.1).

Our approach for implementing data augmentation was to generate new sentences that are equal to the original one, but with one random word that is substituted by a different word, semantically similar to the original one. We utilized established approaches from literature [9, 10] to implement word substitution in the text, by changing the value of the portion of the text and maintaining all the rest unchanged. As an example, the original sentence "Fa ancora na sfaccim e per andare in #moto sulle mie montagne" was transformed into "Fa ancora na sfaccim e per tornare in #moto sulle mie montagne", since the verbs "andare" and "tornare" are semantically similar and do not change the global meaning of the sentence.

In order to find similar words, we used a Word Embedding model for Italian[7] (due to the similarity between Italian and the majority of linguistic varieties spoken in Italy) fine-tuned on our training set, and then we selected one among the vectors that are closer to the vector of the word we want to substitute, using the library "Word2Vec.most_similar"[8].

The augmented dataset ensures an equal distribution of sentences (5549 sentences each) for every region. This quantity corresponds to the initial number of sentences for Lazio, the region with the highest number of entries. In each region, except for Lazio, the number of newly generated sentences equalled the difference between the initial number of sentences in that region and the initial number of sentences in Lazio. Therefore, if a region had a lower initial sentence count, the same sentence was used more frequently for augmentation (thus the number of times a single sentence is used for augmentation is 5549 divided by the number of sentences). As an example, let's consider Sicily which had 608 initial sentences. In this case, each sentence has been used $\approx$ 9 times to create new data, resulting in 4941 newly generated sentences.

3.2. Prediction Through Language Model

The process for classifying the sentences using BERT was the following: (i) we fine-tune BERT on our augmented training set (cfr. § 3.1); (ii) for every sentence in the test set, we use the model to get a prediction on the regional variety of the sentence; (iii) a confidence score for each region is returned.

[5] https://www.vocabolariofiorentino.it/ricerca/lemmi, *retrieved on May 10th, 2023*

[6] https://www.theromanpost.com/2016/06/dizionario-dialetto-romanesco, *retrieved on May 10th, 2023*

[7] https://github.com/MartinoMensio/it_vectors_wiki_spacy

[8] https://tedboy.github.io/nlps/generated/generated/gensim.models.Word2Vec.most_similar.html

Run	Precision	Recall	Macro F1
Standard-Track_Run-1	**0.83**	**0.52**	**0.56**
Standard-Track_Run-2	0.69	0.45	0.48
Standard-Track_Run-3	0.75	0.50	0.52
Logistic_regression baseline	0.62	0.42	0.46
Most_frequent baseline	0.05	0.21	0.07
Special-Track_Run-1	0.72	0.80	0.73
Special-Track_Run-2	0.72	0.80	0.73
Special-Track_Run-3	0.81	**0.83**	**0.82**
Logistic_regression baseline	**0.92**	0.67	0.71
Most_frequent baseline	0.39	0.50	0.44

Table 1
Results of the classification of Italian regions. The upper group shows the performances of the 3 runs for the standard track, compared to the "logistic regression" and "most frequent" baselines. The lower group shows the performances of the 3 runs for the special track, compared with the correspondent baselines.

3.3. Prediction Through Vocabularies

Among the vocabularies presented in § 2.3, we selected the ones relevant to the specific task (it will be discussed in § 4). The first step was to normalize the format of the different vocabularies, associating each regional word with its frequency value. For the vocabularies retrieved from the web, we first performed web-scraping, disregarding the standard Italian translation of the terms, and then assigned 1 as the frequency for each one of the vocabulary entries. Since the vocabularies generated from the training set came with different frequencies, we had to normalize those values, reassigning a value between 0 and 1, by maintaining the same proportion of the original frequencies. A process of normalization was also performed on the vocabulary words: accented characters were converted into unaccented equivalents, IPA representations presented in the terms were deleted, combinations of words (sometimes multiple entries were considered as one in the web vocabulary) were divided into individual entries (e.g., the entry "c(o/u)mpà" becomes two different entries "compa" and "cumpa").
Once we obtained all the normalized vocabularies, we then merged them into one single vocabulary, by summing all frequencies for the same term in the same variety (e.g. if "compà" has frequency 1 in Vocabulary_A for Sicilian and frequency 0.9 in Vocabulary_B, it will have frequency 1.9 in the global vocabulary, assuming there are only 2 vocabularies).
Finally, to predict the regional variety of a sentence, we sum the frequencies of each word in the sentence for each regional variety, using the frequencies present in the global vocabulary. The region with the highest scoring is the predicted regional variety of the sentence.

3.4. Global Prediction

Our final prediction on the region of origin a given sentence leverages both the predictions from BERT (cfr. § 3.2) and from the vocabularies (cfr. § 3.3). We introduce a variable K which regulates the proportions of the contributions given by each one of the 2 predictions. We define the sum of the 2 predictions S for every regional variety r as follows:

$$S(r) = B[r] + K * V[r] \tag{1}$$

where $B[r]$ is the confidence of BERT for the region r and $V[r]$ is the confidence of the vocabularies algorithm for the region r.
The final prediction for the believed variety of the sentence is defined by the following expression:

$$\text{global_pred} = \arg\max_{r \in R} S(r) \tag{2}$$

where R is the set of all regional varieties.

4. Experimental Setup

Our experiments were divided into 2 parts: the first one aims at classifying a regional variety among all the 20 Italian regions and was targeted to the Standard Track of GeoLingIt (cfr. § 4.1), while the second one aims at classifying only Toscana and Lazio varieties and was targeted to the Special Track of GeoLingIt (cfr. § 4.2).

4.1. Standard Track

For classifying one variety among the 20 Italian regions, we used the 3 vocabularies presented in § 2.3 and the BERT classifier, as described in § 3. During the fine-tuning process of BERT, we experimented with different numbers of training epochs, learning rates, and values of K (cfr. § 3.4), and used the 3 configurations that gave

better results on the validation set as the 3 runs of the task:

- **Run-1**: 2 epochs, learning rate $6 * e^{-5}$, K=1;
- **Run-2**: 2 epochs, learning rate $2 * e^{-5}$, K=1;
- **Run-3**: 2 epochs, learning rate $6 * e^{-5}$, K=0.5.

Although we tried with a lower and greater number of epochs, we observed that 2 was the best value for the relatively small training set that we used for fine-tuning. The same approach was used to set the Adam optimiser's learning rate. We started with a learning rate of $2 * e^{-5}$ as suggested by TensorFlow documentation, and gradually decrease or increase it.
Finally, we noted that relying too much on the vocabularies rather than the LLM (using a K greater than 1), did not bring high results, thus we focused on values of K between 0 and 1.

4.2. Special Track

We followed the same process also for the special track, using BERT fine-tuned on a corpus of only Toscana and Lazio samples (filtered out from the augmented dataset, described in § 3.1), and the 4 vocabularies presented in § 2.3, 2 generated from the original training-set, one for the Toscana lexicon, and one for the Lazio lexicon. We tried different configurations and used the best 3 as the runs for the special track:

- **Run-1**: 2 epochs, learning rate $4 * e^{-5}$, K=0.5;
- **Run-2**: 2 epochs, learning rate $4 * e^{-5}$, K=0.1;
- **Run-3**: 2 epochs, learning rate $2 * e^{-5}$, K=0.1.

Again, we found that 2 epochs were the optimal value on the validation set and that the best learning rate values were around the suggested value from the literature. In this case, we observed that the contribution of the vocabularies did not bring many advantages, and therefore we kept a low value of K.

5. Results

Table 1 shows the results for our 3 runs for the standard track (classification on the 20 Italian regions) and 3 runs for the special track (classification on the Toscana and Lazio varieties), as described in section 4. In addition, we reported also the results for the 2 baselines provided by the organizers of the tasks, one based on a logistic regression model and the other one which simply predicts the most frequent label in the training set for every inference, as described by Ramponi and Casula [11]. These baselines are reported in the table once per group and are relative to the task of the correspondent group.
According to the task's indications, we employ macro-averaged precision, recall and f1-score as evaluation metrics.

6. Discussion

For the standard track group, we can observe that our best run obtained a value for the f1-score 8 times higher than the Most_frequent baseline and 10 points higher than the Logistic_regression baseline. Comparing the 3 runs, besides the differences in the hyper-parameters choice (a learning rate of $6 * e^{-5}$ appears to give better results), it is interesting to note that the use of the vocabulary has a positive effect on the performances, since we observe an improvement in all evaluation metrics when we pass from $K = 0.5$ to $K = 1$ (from Run-3 to Run-1, which are identical for all other parameters).
For the special track group, we improved the Most_frequent and Logistic_regression baselines of 38 and 11 points respectively, for what concerns the f1-score. Here, in contrast, the use of the vocabulary seems to be less influential than the learning rate, which results to work better with a value of $6 * e^{-5}$.
This difference between the two tasks can be partially explained by the relative lexical similarity between Toscana and Lazio varieties: in both regions, the regional spoken language does not differ too much from standard Italian and therefore the strategy of distinguishing a sentence on the basis of the vocabulary does not seem to be the right approach. On the other side, when it comes to classifying a sentence between all the 20 Italian regions varieties, the lexical terms differ significantly and therefore our prediction based on the vocabularies proves to be a great improvement to the classification made through LLMs only.

7. Conclusion

In this paper, we addressed the classification of *non-standard* Italian Twitter posts according to their regional variety, by combining the prediction obtained using a language model (BERT) with an algorithm utilizing regional varieties dictionaries and word frequency. We contribute to two tasks: a classification at the country level, and a classification according to a subset of regions (Lazio and Toscana).

After briefly introducing the Italian regional varieties, the tasks addressed, and the methodology applied, a description of the data provided and other additional resources retrieved (e.g., online vocabularies) follows. We then explain in detail the methodology used, starting from the augmentation of the training set data, and going through the different techniques used for obtaining the intermediate and final predictions for both tasks. Finally, the evaluation results are shown: the first task achieved a macro F1 score of 0.56, outperforming both the logistic regression baseline (0.46) and the most frequent baseline

(0.07); while in the second task, even if the macro F1 score and the recall show substantial improvement, the overall precision is 11 points lower with respect to the logistic regression baseline (0.81 vs 0.92).

Overall, the knowledge captured from regional dictionaries and word frequencies seems to be effective in capturing the nuances and characteristics of regional varieties. Furthermore, by leveraging BERT's bidirectional pre-training, the system can consider the entire context of a sentence, thereby contributing to accurate predictions.

The availability of curated resources, such as regional dictionaries, played an important role in enhancing the system's performance. However, we acknowledge the limitations of the vocabularies used in our experiments, and further efforts should be made to expand and refine these resources. Finally, we think there are still areas for improvement. Future research could explore more sophisticated and tailored methods for data augmentation and investigate alternative techniques for integrating vocabulary lexical analysis with BERT. Moreover, the inclusion of additional linguistic features and the exploration of ensemble methods could potentially lead to further performance improvements.

References

[1] C. Moseley, Atlas of the World's Languages in Danger, Memory of peoples Series, UNESCO Publishing, 2010. URL: https://books.google.it/books?id=kFVthqmDs_kC.

[2] C. Bosco, F. Dell'Orletta, S. Montemagni, M. Sanguinetti, M. Simi, The evalita 2014 dependency parsing task, The Evalita 2014 Dependency Parsing Task (2014) 1–8.

[3] F. Dell'Orletta, Ensemble system for part-of-speech tagging, Proceedings of EVALITA 9 (2009) 1–8.

[4] M. Polignano, P. Basile, M. De Gemmis, G. Semeraro, V. Basile, et al., Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets, in: CEUR Workshop Proceedings, volume 2481, CEUR, 2019, pp. 1–6.

[5] A. Ramponi, Nlp for language varieties of italy: Challenges and the path forward, arXiv preprint arXiv:2209.09757 (2022).

[6] A. Ramponi, C. Casula, GeoLingIt at EVALITA 2023: Overview of the geolocation of linguistic variation in Italy task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[7] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[8] J. Devlin, M.-W. Chang, K. Lee, Google, kt, language, ai: Bert: pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT, 2019, pp. 4171–4186.

[9] W. Wang, Z. Zhang, J. Guo, Y. Dai, B. Chen, W. Luo, Task-oriented dialogue system as natural language generation, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 2698–2703.

[10] T. Labruna, B. Magnini, Fine-tuning bert for generative dialogue domain adaptation, in: Text, Speech, and Dialogue: 25th International Conference, TSD 2022, Brno, Czech Republic, September 6–9, 2022, Proceedings, Springer, 2022, pp. 513–524.

[11] A. Ramponi, C. Casula, Diatopit: A corpus of social media posts for the study of diatopic language variation in italy, in: Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023), 2023, pp. 187–199.

baρtti at GeoLingIt: Beyond Boundaries, Enhancing Geolocation Prediction and Dialect Classification on Social Media in Italy

Alkis Koudounas[1], Flavio Giobergia[1], Irene Benedetto[1,2], Simone Monaco[1], Luca Cagliero[1], Daniele Apiletti[1] and Elena Baralis[1]

[1]Department of Control and Computer Engineering, Politecnico di Torino, Turin, Italy
[2]MAIZE, Turin, Italy

Abstract
The proliferation of social media platforms has presented researchers with valuable avenues to examine language usage within diverse sociolinguistic frameworks. Italy, renowned for its rich linguistic diversity, provides a distinctive context for exploring diatopic variation, encompassing regional languages, dialects, and variations of Standard Italian. This paper presents our contributions to the GeoLingIt shared task, focusing on predicting the locations of social media posts in Italy based on linguistic content. For Task A, we propose a novel approach, combining data augmentation and contrastive learning, that outperforms the baseline in region prediction. For Task B, we introduce a joint multi-task learning approach leveraging the synergies with Task A and incorporate a post-processing rectification module for improved geolocation accuracy, surpassing the baseline and achieving first place in the competition.

Keywords
natural language processing, dialect localization, diatopic variation, deep learning

1. Introduction

The advent of social media has significantly facilitated the investigation of language usage across diverse sociolinguistic aspects. Italy, in particular, stands out as a compelling case study due to its remarkable diatopic variation, encompassing an array of local languages, dialects, and regional manifestations of Standard Italian within a relatively confined geographic area [1]. This linguistic heterogeneity stems from historical and cultural influences, with distinct lexicons, grammatical structures, and pronunciations shaping the various language varieties present in the country, each bearing the imprints of historical events, geographical isolation, and cultural traditions. Furthermore, the integration of regional varieties of Standard Italian further enriches the linguistic mosaic of Italy [2]. Within the digital realm, particularly on platforms like Twitter, Italian speakers leverage these linguistic variations to express their social identities and affiliations, thereby contributing to the visibility and preservation of these diverse linguistic forms in the online domain. This intriguing sociolinguistic phenomenon has attracted researchers from computational linguistics and sociolinguistics domains, providing valuable insights into the nuances of language variation in Italy.

The GeoLingIt shared task [3] at Evalita 2023 [4] aims to advance the current knowledge of linguistic variation in Italy by focusing on the prediction of locations of social media posts from Twitter based solely on linguistic content. In this paper, we present our contributions to the GeoLingIt shared task. GeoLingIt proposes two separate tasks: Subtask A, a classification task that aims to identify the region of provenance of a tweet exhibiting non-Standard Italian language, and Subtask B, a regression task to identify the fine-grained location of the provenance of the same tweets, in terms of longitude and latitude coordinates. Both tasks are based on the DiatopIt dataset [5].

For Task A, we first gather additional data sources specifically for the various Italian regions. We propose a novel approach involving a pre-training step of state-of-the-art transformer-based models with a contrastive learning strategy leveraging data augmentation techniques. This approach outperforms the baseline, demonstrating the effectiveness of leveraging pre-training and contrastive learning to improve the accuracy of region prediction. For Task B, we introduce a joint multi-task learning approach that addresses the challenge of fine-grained variety geolocation. Our approach outperforms the baseline by simultaneously tackling both tasks. Additionally, we introduce a post-processing rectification module that refines the predicted coordinates and ensures their alignment within the boundaries of Italy. This module enhances the reliability of the predicted locations, making them more precise and geographically accurate.

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
alkis.koudounas@polito.it (A. Koudounas)
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper16.pdf

Our proposed methods[1] not only achieve state-of-the-art performance in terms of location prediction (with a test error lower than 100 km compared to an error of more than 250km of the baseline methods) but also offer some valuable insights into Italy's diverse linguistic landscape.

2. Related work

The analysis of linguistic varieties and dialects is an emerging topic in the field of Natural Language Processing (NLP) [6].

Efforts have been made to address and incorporate variations in corpora, such as pronunciation and spelling differences.

Gelly et al. [7] and Elfeky et al. [8] addressed the language varieties for speech recognition of the English language, emphasizing the fact that the task is even more challenging as the space of dialects is broad. Many researchers in the past have attempted to address the challenge of language variation by leveraging social media data. Grieve et al. [9] compared regional patterns, both analyzing dialect labels and geolocation, finding strong correlations between the two sources. Sadat et al. [10] have shown that probabilistic models of language identification can be used to identify Arabic dialects on tweets. Efforts in the direction of propagating information (e.g., sentiment) from high-resource languages (e.g., Italian) to low-resource ones (e.g., regional variations) through vector space alignments have shown promising results, as shown by Giobergia et al. [11] across languages (from English to other ones). Recently, Italian computational linguistics research has encountered difficulties due to the limited availability of large-scale datasets specifically tailored for the language, as emphasized in a recent study [12]. Unfortunately, the substantial computational resources needed for pre-training language models have resulted in only a few architectures being accessible for Italian.

Moving to the geolocation task, Han et al. [13] proposed a method for geolocation prediction based on identifying location-indicative words. Nevertheless, the work was not focused on dialects. Eisenstein et al. [14] inspected the correlation between geographical information and sociolinguistic associations instead of predicting the demographical attributes of users based on their tweets and their position. Other works have focused on the geolocation task [15, 16], but they do not take into account the language varieties.

[1]The code to reproduce all our experiments is available at https://github.com/koudounasalkis/barotti-GeoLingIt2023

3. Methodology

This section outlines the methodology adopted to automatically ascertain the region (**Task A**) and the coordinates, in terms of latitude and longitude (**Task B**), of the origin of a tweet, considering the pronounced class imbalance prevalent within the training dataset. To tackle these challenges, we propose to use two key techniques: data augmentation and pre-training with contrastive learning, and multi-task learning.

3.1. Task A

The goal of task A is to identify the origin region of a given tweet. We denote the set of regions as R. Data augmentation plays a crucial role in our methodology, as it is implemented to address the class imbalance during the training process. In the initial data collection phase, we obtain a substantial amount of regional dialect data from various online sources[2]. Dump, editions, and further information on the collected data are available in the official repository. We then pre-process them, obtaining an expanded vocabulary that is utilized for the purpose of data augmentation. We denote the vocabulary for each region r as $\mathcal{D}_r$ ($r \in R$).

We adopt a substitution approach to words in tweets representing language variations to build an augmented version of the original dataset. Each tweet $x_i = \{t_{1,i}, ..., t_{M,i}\}$ belonging to region y_i is augmented by randomly replacing words that are contained in $\mathcal{D}_{y_i}$ with other words from the same region, with a random probability p. More formally, each term $t_{i,k} \in x_i$ is replaced with $t'_{i,k}$, defined follows:

$$t'_{i,k} = \begin{cases} t \sim \mathcal{D}_{y_i} & \text{if } p \leq p' \\ t_{i,k} & \text{otherwise} \end{cases} \quad (1)$$

Where p' is the probability of replacing each term $t_{i,k}$ with a different one $t'_{i,k}$ drawn from the same region $\sim \mathcal{D}_{y_i}$. We experimentally observe the best results in terms of performance for $p' = 0.5$.

Regarding the contrastive learning strategy, we pre-train the model to enhance its ability to discern whether two tweets belong to the same region. During this preliminary training phase, the model learns to differentiate between tweet pairs and their corresponding regional affiliations. Given two tweets, denoted as x_i and x_j, along with their labels y_i and y_j, the model is trained to minimize a loss that facilitates this discrimination:

$$\mathcal{L}_{contr} = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1,k\neq i}^{2N} \exp \text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau} \quad (2)$$

[2]Italian Wikipedia: it.wikipedia.org, Dialettando: dialettando.com [accessed May-2023]

where $\mathbf{z}_i$ and $\mathbf{z}_j$ are the latent representations learned by the model of the tweets x_i and x_j. We set the temperature parameter τ equal to 1, and the function `sim` is the cosine similarity. In this approach, we randomly select a sample from the dataset to serve as an anchor. We then create a positive data point by augmenting the anchor with words from the same region and a negative sample by augmenting the anchor with words from a different region.

We call this approach *Contrastive PT & Data++* (See section 5 for more details). Pre-training with contrastive learning and data augmentation techniques is specifically devised to mitigate the challenges posed by the imbalanced class distribution within the dataset. This has been proven beneficial in several tasks and domains [17, 18, 19].

Additionally, we empirically observed that a simple logistic regression model performs well in terms of F1 score on various minority classes. We attribute this to a lower model capacity, reducing the amount of overfitting that may occur in minority classes. To leverage this insight, we propose using an exclusive class assignment mechanism (named *Entropy-based Ensemble* in the following) that uses the confidence of the BERT-based model (further information about the model in Section 4). In other words, when the BERT-based prediction is made with low confidence (according to a specific empiric threshold), we replace the overall prediction with the one made by the logistic regression if the latter's confidence is higher.

We estimate the confidence of the BERT-based model using the entropy of its predicted probabilities. Lower entropy is associated with high certainty (i.e., the model predicts one class with high probability and all others with a low one), and vice-versa.

For the training of the logistic regression, we use as input, for each tweet, the respective bag of words as well as the vector $d_i = \{|t_i \cap \mathcal{D}_r|, r \in R\}$, i.e., the number of words within each tweet that belong to each region in our dictionary.

3.2. Task B

The objective of this task is to automatically determine the geographical coordinates (latitude and longitude) of the origin of a tweet. Recognizing the strong correlation between Task A and Task B, we adopt a multi-task learning approach to tackle them jointly.

In the multi-task learning setup (the *Multi-Task FT* approach), we build a two-head model by adding to the classification a regression layer, enabling the estimation of coordinates together with the regional classes. The model is thus trained to simultaneously learn both the geographical location and the class of the tweet. To achieve this, we optimize the model using a weighted combination of loss functions. For Task A, we aim to maximize the F1 score by minimizing the corresponding cross-entropy loss. For Task B, we minimize the Haversine distance by minimizing the mean squared error (MSE) loss, which helps to reduce the difference between the predicted and target coordinates. Given a tweet, denoted as x_i, the model estimates a loss function $\mathcal{L}_{CR}$ that encompasses both tasks, minimizing the weighted conjunction of a standard cross-entropy loss for classification $\mathcal{L}_C$ and an L-2 loss for regression $\mathcal{L}_R$:

$$\mathcal{L}_{CR} = \mathcal{L}_C + \alpha \mathcal{L}_R \tag{3}$$

where:

$$\mathcal{L}_C = \mathcal{L}(\hat{y}^c, y^c) = -\sum_{i=1}^{N} y_i^c \log(\hat{y}_i^c) \tag{4}$$

$$\begin{aligned} \mathcal{L}_R &= \mathcal{L}(\hat{y}^{lat}, y^{lat}) + \mathcal{L}(\hat{y}^{lon}, y^{lon}) \\ &= \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i^{lat} - y_i^{lat})^2 + \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i^{lon} - y_i^{lon})^2 \end{aligned} \tag{5}$$

where y^c is the classification label, i.e., the region, y^{lat}, y^{lon} are the latitude and the longitude respectively, N is the number of samples, and α represents the weight assigned to the regression loss once it has been adjusted to have a similar magnitude as the classification one. The optimal value was determined to be $\alpha = 0.5$.

Please note that this approach does not leverage the *Contrastive PT & Data++* pre-training strategy.

In addition to the joint task learning, we introduce a rectification module to refine the model's predictions (we denote this model *"Beyond-Boundaries" Multi-Task FT*). This module leverages the geographical domain knowledge that dictates that tweets are expected to be found within the territorial confines of the country. In other words, it ensures that tweets are geographically located within the national borders, specifically on land rather than in the sea. Coordinates that fall outside of these constraints are adjusted so as to be set to the closest point within the boundaries. By incorporating these techniques, we aim to enhance the model's performance in both identifying the geographical origin of a tweet and classifying its region. The multi-task learning framework enables us to leverage the interdependence between the two tasks. At the same time, the rectification module ensures that the predicted coordinates remain within the boundaries of Italy. We enforce this constraint as a post-processing step, where coordinates are projected onto a high-resolution map of Italy (with a granularity of 1.5 km). Points that fall outside of the boundaries of Italy are moved to the closest point within the country.

For the sake of completeness, we conclude the analysis with an approach that merges the pre-training with contrastive learning and data augmentation strategy with the multi-task fine-tuning scheme. We call this approach *"Continuous Learning"*.

Method	Val F1	Test F1
Most frequent baseline	2.65%	7.38%
Logistic regression	58.72%	46.11%
Contrastive PT & Data++	**72.61%**	**53.18%**
Multi-Task FT	55.25%	51.72%
Entropy-based Ensemble	68.78%	51.74%

Table 1
Results (%) on dev and test sets for **task A**. Best results are highlighted in bold. First two rows are the baselines given by the task organizers.

Method	**Val Dist**	**Test Dist**
Centroid baseline	301.65	281.04
k-nearest neighbors	281.03	263.35
Multi-Task FT	111.05	120.02
"Continuous Learning" Multi-Task FT	99.50	98.79
"Beyond Boundaries" Multi-Task FT	**98.41**	**97.74**

Table 2
Results (in km) on dev and test sets for **task B**, computed as the Haversine distance. Best results are highlighted in bold. First two rows are the baselines given by the task organizers.

4. Experimental setting

Models. We consider various models, including Italian-BERT [20] *cased* and *uncased* versions, LABSE [21] and BART-IT [22] models pre-trained for the Italian language. We find the best model, based on the performance obtained on the validation set, to be BERT-BASE-ITALIAN-UNCASED. Thus this is the base model used to address the tasks. All the pre-trained checkpoints of these models are taken from the Hugging Face hub repository[3].

Hyperparameter Setup. We ran a manual hyperparameter search and followed fine-tuning procedures and guidelines from relevant literature. We provide detailed information about the models used for the evaluation, the hyperparameter setup, and the fine-tuning procedure in the official project repository.

5. Results

Table 1 presents results for task A, evaluating different methods based on their validation and test F1 macro scores. The proposed novel approach, which combines contrastive pre-training and data augmentation, demonstrated superior performance compared to other methods. It achieved the highest F1 scores on the validation and test sets, reaching 72.61% and 53.18%, respectively. We believe that the reason for this performance difference lies in the presence of new out-of-distribution samples in the test set, which our model struggles to recognize accurately. Interestingly, while the proposed multi-task model excels in task B (*Multi-Task FT*), surpassing baseline models performance as shown later, it fails to deliver satisfactory results for task A. Nonetheless, it still outperforms the baseline. Conversely, the *Entropy-based Ensemble* method, which enhances BERT performance with LR's one, achieves a high score on the validation set. However, it only slightly outperforms the *Multi-Task* approach on the test set, with an improvement of 0.02%.

In Table 2, various methods are evaluated for task B, based on the Haversine distance metric. The *Multi-Task* model approach demonstrated substantial enhancements compared to the baselines, achieving a validation distance of 111.05 km and a test distance of 120.02 km. This emphasizes that incorporating a multi-objective function helps the model better tackle the given task. Moreover, training the model in a multi-task manner, starting from the pre-trained model that underwent contrastive learning and data augmentation and was fine-tuned for task A (referred to as *"Continuous Learning"* in Table 2), resulted in a significant performance boost (test distance of 98.79 km). This improvement is likely attributed to the model already possessing domain knowledge, leading to improved performance on the test set. Lastly, the additional rectification *"Beyond-Boundaries"* module effectively refines the precision of the model achieving the best performance on the test set (97.74 km) and securing the first position in the GeoLingIt challenge.

5.1. Logistic regression insights

As discussed, we used logistic regression in combination with the BERT-based solution to override low-confidence predictions. The weights learned by the logistic regression can be interpreted as the importance the model assigns to each feature's presence (and magnitude). The features passed are either the number of occurrences of various words or the overall number of words contained that are known to belong to various regions (dialects). Table 3 shows, for each region, the ten features with the largest weights[4].

In some cases (underlined in the table), actual names of regions and cities are also relevant indicators of a tweet's origin. This is a reasonable result, as tweets made in a certain dialect are intuitively likely to mention geographic places related to the dialect itself. We note that the features containing the counts of the words that belong to the various language varieties are also sometimes considered useful indicators by the logistic regression. In

[3] https://huggingface.co/dbmdz/bert-base-italian-uncased latest access: May 2023

[4] Some words may have a negative, offensive, or misogynistic connotation. We still report these results for the sake of thoroughness.

Region	Top 10 features
Abruzzo	fregna, diaco, st, statt, <# **tokens molise**>, <# **tokens puglia**>, asin, cussu, abruzzo, ju
Basilicata	fandom, mezzarella, accusci, ah, pazz, aggia, cazz, hahahaha, cazzu, trmon
Calabria	<# **tokens calabria**>, calabria, aru, nto, ccu, ciota, capu, jamu, frica, fimmina
Campania	napoli, foss, semp, merd, statt, strunzat, ua, ata, sciem, lota
Emilia Romagna	socmel, maial, tin, sempar, bologna, cinno, umarell, cagher, veh, soccia
Friuli-Venezia Giulia	femo, triestin, magnar, ocio, ga, gavemo, mona, trieste, xe, orpo
Lazio	avoja, annamo, avemo, mortacci, artra, nse, artro, aspettamo, ar, stamo
Liguria	ou, cusci, abbelinati, porcu, rasciun, emma, pestu, semmu, zena, zeneize
Lombardia	dighel, pirlata, pheega, nanca, danee, milano, gh, sciuri, gnaro, sciur
Marche	en, sperem, ecche, roscio, ancona, marche, scritturebrevi, daje, diaulu, sblab2021
Molise	pipponi, ah, buongiornoatutti, venta, fior, fatt, vientu, paes, sort, fake
Piemonte	picio, piou, boja, piemonte, suma, speruma, fauss, nen, piciu, babaciu
Puglia	salentu, capu, mang, isolitiignoti, munnu, mme, trimone, arret, trmon, bari
Sardegna	ajo, macca, sardegna, <# **tokens sardegna**>, tottu, biri, tontu, nudda, sesi, itte
Sicilia	chidda, quantu, nuddu, camurria, fici, soddi, carusi, bonu, semu, <# **tokens sicilia**>
Toscana	guasi, nsomma, <# **tokens toscana**>, caa, siuro, boja, diaccio, oglioni, gnamo, tope
Trentino-Alto Adige	10, bicer, maial, tasi, ghe, sberloni, stinc, sior, tai, pu
Umbria	pija, ch, <# **tokens umbria**>, er, porchetto, mejo, bbona, mixatino, je, umbria
Valle d'Aosta	carbonada, buonissimo, int, piacione, nasconderti, max, bosc, devise, cher, vivre
Veneto	varda, sboro, dixe, queo, casin, venessia, ciava, <# **tokens veneto**>, xe, veneto

Table 3
Top-10 relevant features identified by the logistic regression for each region, based on the magnitude of the weights learned. Underlined are the tokens that refer directly to a region or city (in their Italian form). In **bold** the feature related to the number of words of the respective region.

most cases, the count used is the relevant one for the region of interest (for example, the number of tokens from the Venetian language varieties, <# tokens veneto> is a valuable feature to detect the "Veneto" region). The only exception occurs for Abruzzo, where the presence of both token counts from Molise and Puglia are considered helpful indicators. Given the geographic proximity of these regions, we find this result to be reasonable.

Finally, it can be observed that some situations arise where words that are generally not characterizing for certain regions still emerge as being significant ones (e.g., "hahahaha" for Basilicata, or "sblab2021" for Marche). We believe this to be an overfitting problem due to the lack of meaningful data on some of the minority regions: as such, it could be addressed by collecting additional data for those regions.

6. Conclusion and future work

This paper presented our contributions to the GeoLingIt shared task. We addressed Task A by designing a pretraining approach that leverages data augmentation and contrastive learning, surpassing the baseline and demonstrating the effectiveness of our approach in region prediction. For Task B, we introduced a joint multi-task learning approach that outperformed the baseline and incorporated a post-processing rectification module, resulting in precise and geographically accurate location predictions. Our methods not only achieved state-of-the-art performance, allowing us to be placed first for Task B, but also provided some model insights into the rich linguistic landscape of Italy.

Future work could delve into fine-grained dialect classification. This involves developing models capable of identifying specific dialects or regional varieties within a given region, which would provide a more nuanced understanding of language variation in Italy and enable more targeted analyses of sociolinguistic phenomena.

Acknowledgments

This study was carried out within the FAIR - Future Artificial Intelligence Research and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013), the grant "National Centre for HPC, Big Data and Quantum Computing", CN000013 (approved under the M42C Call for Proposals - Investment 1.4 - Notice "Centri Nazionali" - D.D. No. 3138, 16.12.2021, admitted for funding by MUR Decree No. 1031,17.06.2022), as a part of the MALTO (MAchine Learning @ poliTO) team, with partial support from SmartData@PoliTO center on Big Data and Data Science. This

manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

References

[1] M. Maiden, M. Parry, The dialects of Italy, Routledge, 2006.

[2] A. Ramponi, Nlp for language varieties of italy: Challenges and the path forward, arXiv preprint arXiv:2209.09757 (2022).

[3] A. Ramponi, C. Casula, GeoLingIt at EVALITA 2023: Overview of the geolocation of linguistic variation in Italy task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[4] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[5] A. Ramponi, C. Casula, DiatopIt: A corpus of social media posts for the study of diatopic language variation in Italy, in: Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023), Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 187–199. URL: https://aclanthology.org/2023.vardial-1.19.

[6] M. Zampieri, P. Nakov, Y. Scherrer, Natural language processing for similar languages, varieties, and dialects: A survey, Natural Language Engineering 26 (2020) 595–612.

[7] G. Gelly, J.-L. Gauvain, L. Lamel, A. Laurent, V. B. Le, A. Messaoudi, Language recognition for dialects and closely related languages., in: Odyssey, volume 2016, 2016, pp. 124–131.

[8] M. G. Elfeky, P. Moreno, V. Soto, Multi-dialectical languages effect on speech recognition: Too much choice can hurt, Procedia Computer Science 128 (2018) 1–8.

[9] J. Grieve, C. Montgomery, A. Nini, A. Murakami, D. Guo, Mapping lexical dialect variation in british english using twitter, Frontiers in Artificial Intelligence 2 (2019) 11.

[10] F. Sadat, F. Kazemi, A. Farzindar, Automatic identification of arabic language varieties and dialects in social media, in: Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP), 2014, pp. 22–27.

[11] F. Giobergia, L. Cagliero, P. Garza, E. Baralis, Cross-lingual propagation of sentiment information based on bilingual vector space alignment., in: EDBT/ICDT Workshops, 2020, pp. 8–10.

[12] A. Koudounas, M. La Quatra, L. Vaiani, L. Colomba, G. Attanasio, E. Pastor, L. Cagliero, E. Baralis, Italic: An italian intent classification dataset, arXiv preprint arXiv:2306.08502 (2023).

[13] B. Han, P. Cook, T. Baldwin, Geolocation prediction in social media data by finding location indicative words, in: Proceedings of COLING 2012, 2012, pp. 1045–1062.

[14] J. Eisenstein, N. A. Smith, E. Xing, Discovering sociolinguistic associations with structured sparsity, in: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, 2011, pp. 1365–1374.

[15] K. Lee, R. Ganti, M. Srivatsa, L. Liu, When twitter meets foursquare: tweet location prediction using foursquare, in: 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, 2014.

[16] A. Rahimi, T. Cohn, T. Baldwin, Twitter user geolocation using a unified text and network prediction model, arXiv preprint arXiv:1506.08259 (2015).

[17] X. Wang, G.-J. Qi, Contrastive learning with stronger augmentations, IEEE Transactions on Pattern Analysis and Machine Intelligence (2022).

[18] R. Zhang, Y. Ji, Y. Zhang, R. J. Passonneau, Contrastive data and learning for natural language processing, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts, 2022, pp. 39–47.

[19] L. Vaiani, A. Koudounas, M. La Quatra, L. Cagliero, P. Garza, E. Baralis, Transformer-based non-verbal emotion recognition: Exploring model portability across speakers' genders, in: Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge, 2022, pp. 89–94.

[20] S. Schweter, Italian bert and electra models, 2020. URL: https://doi.org/10.5281/zenodo.4263142. doi:10.5281/zenodo.4263142.

[21] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, W. Wang, Language-agnostic bert sentence embedding, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 878–891.

[22] M. La Quatra, L. Cagliero, Bart-it: An efficient sequence-to-sequence model for italian text summarization, Future Internet 15 (2022) 15.

Salogni at GeoLingIt: Geolocalization by Fine-tuning BERT

Ilaria Salogni[1]

[1]*Università di Pisa, Italy*

Abstract

The recent growing interest in low-resource languages has been significantly bolstered by transformer-based models. By fine-tuning three such models, two based on BERT and the other on RoBERTa, I aim at geolocating sequences exhibiting non-standard language varieties relying solely on linguistic content. I find that, given that the information contained in the embeddings is all we need to carry out this complex task, a model architecture with less task-specific layers leads to better results. Furthermore, models pre-trained on miscellaneous corpora generalize better than those trained exclusively on tweets. The work also shows that the greater availability of resources of a certain regional variety positively affects the capacity of the model.

1. Introduction

Recognizing varieties and forming an opinion about where the speaker comes from is something so ingrained in our experience as speakers that it seems innate, and even a little magical. The question that drives this work is: can Large Language Models (LLMs) do what we do and if so, how well can they do it? do they do that in a way that is operationally similar to ours? The Italian scenario is a good testing ground as despite his limited geographical extent, it is one of the most linguistically-diverse in Europe. In their work, Ramponi and Casula say that *current transformer-based models are rather limited for modeling language variation over space in highly multilingual areas such as Italy* [1]. I don't agree completely, not only because of the encouraging results of the application of LLMs in a always growing number of tasks, but also because what we can explain on how they work does not highlight anything which may prevent good performance. Furthermore, the work of Lutsai and Lampert [2] reaches the astonishing result of a median error of 30km worldwide level, and fewer than 15 km on the US-level datasets for the models trained and evaluated on text features of tweets' content and meta data context, using a BERT model [3]. The fact that Twitter language identifier classifies with the label designed for standard Italian language also contents both partially and fully written in language varieties of Italy, as observed again by Ramponi and Casula [1] may suggest that the LLMs already have in their pre-training dataset the knowledge that they need to carry out a geolocalization task.

This document describes the model I submitted to the EVALITA 2023 evaluation campaign [4] for the task GeoLingIT [5].

1.1. Task

The goal of this project is to predict its location in terms of longitude and latitude coordinates (fine-grain geolocation) of tweets exhibiting non-standard language, based solely on linguistic content. This is a (double) regression task. In contrast to previous geolocation shared tasks on other areas ([6]; [7]; [8]), GeoLingIt is focused on Italy.

1.2. Dataset

GeoLingIT task data comprises 15K geotagged tweets that exhibit non-standard Italian language use (the content may be fully written in local language varieties or exhibiting code-switching with standard italian), and that have been collected in the corpus DiatopIt [1]. The data is annotated with latitude and longitude. After removing the emojis and tags, all the labeled data provided by the organizers were merged and then split into train-eval-test sets. Several crossvalidations were performed with 3-folds or 2-folds split, using train-eval sets. Target and output coordinate data were normalized using Min-Max scaling, as this understandably improved the quality of model prediction.

2. System description

Knowing that representations learned by transformer-based models achieve strong performance across many tasks with various datasets ([9], *inter alia*), I first decided to perform the fine-tuning of three different monolingual BERT-based [3] or RoBERTa-based [10] models, pre-trained on Italian texts. After picking the best performing model, I cross-validated it on a diverse set of hyperparameter configurations (e.g., number and size of hidden layers, activation functions) to pick the best task-specific architecture. All the runs were performed on Colab using high-RAM Nvidia A100 GPUs.

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT

i.salogni@studenti.unipi.it (I. Salogni)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper17.pdf

Table 1
averaged MAE on the eval set for each model; batch size 50-100-150; 10 epochs; in bold the best result

	umberto-commoncrawl-cased-v1				bert-base-italian-cased				AlBERTo-it			
	fold1	fold2	fold3	avg	fold1	fold2	fold3	avg	fold1	fold2	fold3	avg
50	0.0147	0.0136	0.0144	**0.0142**	0.0184	0.0179	0.0184	0.0182	0.0216	0.0209	0.021	0.0211
100	0.0159	0.018	0.0176	0.0171	0.0176	0.0206	0.0191	0.0191	0.0232	0.0228	0.0235	0.0231
150	0.0168	0.0163	0.0173	0.0168	0.0181	0.019	0.0202	0.0191	0.0248	0.0239	0.0247	0.0244

Table 2
averaged MAE on the eval set; umberto-commoncrawl-cased-v1 in each architecture and configuration; batch size 50; 10 epochs; in bold the 4 best results, in red the 4 worse results

	No Hidden Layer				1 Hidden Layer (5)				1 Hidden Layer (300)		
	fold1	fold2	fold3	avg	fold1	fold2	fold3	avg	fold1	fold2	avg
Identity	0.0147	0.0136	0.0144	**0.0142**	0.0485	0.0179	0.0282	0.0315	0.0150	0.0145	**0.0147**
Sigmoid	/	/	/	/	0.0174	0.0138	0.0142	**0.0151**	0.0187	0.0151	0.0169
ReLU	/	/	/	/	0.2957	0.0143	0.1462	0.1520	0.0147	0.0152	**0.0149**
	3 Hidden Layers (5, 5, 10)				3 Hidden Layers (10, 5, 5)				3 Hidden Layers (300, 100, 100)		
	fold1	fold2	fold3	avg	fold1	fold2	fold3	avg	fold1	fold2	avg
Identity	0.0434	0.0517	0.0761	0.0570	0.024	0.0204	0.0412	0.0285	0.0228	0.0207	0.0217
Sigmoid	0.0347	0.0275	0.0395	0.0339	0.0229	0.0351	0.0272	0.0284	0.0144	0.0171	0.0157
ReLU	0.1819	0.0303	0.034	0.0820	0.3067	0.1171	0.1822	0.202	0.0784	0.1479	0.1131

2.1. Comparing different models

To assess how different pre-trained monolingual models perform on the given dataset, I fine-tuned the RoBERTa-based umberto-commoncrawl-cased-v1 [11] model, and the BERT-based models bert-base-italian-cased [12] and AlBERTo-it [13], adding a single linear layer with two output neurons to the pooling layer of each model, without activation function. I tested this "minimal" task-specific architecture on 3 batch values (50, 100, 150) for 10 epochs, dividing the train-dev set into 3 folds.

2.2. Adding a single hidden layer

To explore the potential benefits of introducing additional complexity to the model, I designed a new task-specific architecture adding a single hidden layer right after the pooling layer, testing different sizes (5 neurons and then 300 neurons) followed by an activation function (Identity, Sigmoid or ReLU), and finally, a two-neuron output layer. To reduce the computational cost, only umberto-commoncrawl-cased-v1 was tested using this and the next architectures. For the same reason batch size 50 was maintained.

2.3. Adding more hidden layers

Still with the rationale of knowing whether adding further complexity would enhance the model's learning capacity, I tested a task-specific setting with 3 hidden layers with neurons in combination (5, 5, 10) (10, 5, 5) and (300, 100, 100) on various activation functions (Identity, Sigmoid, ReLU) added before the two-neuron output layer.

3. Results

The umberto-commoncrawl-cased-v1 model on the "minimal" task specific architecture yielded the best Mean Absolut Error (MAE) results in 3-fold cross-validation using the provided labelled data, and achieved 128.19 km of avg distance in km on the blind test set provided by the challenge organizer. Although AlBERTo-it and bert-base-italian-cased were outperformed, their achieved results are not too distant, as shown in Table 1.

The second best MAE results were achieved using the 300 neurons single hidden layer task-specific architecture, and then followed by the 5 neurons single hidden layer architecture, as shown in Table 2. This can be explained by thinking that adding a small hidden layer after the pooling layer leads to an initial drastic reduction in the size of the model output.

The worst results, on the other hand, were all obtained with the 3 hidden layers architecture and ReLU as activation function. The accuracy dropped possibly because of excessive feature compression: when several hidden layers are stacked, this reduction is followed by another further reduction of the size of the input vector, and the linear activation function was of no use in this case. Therefore, further complicating the architecture requires an additional regularization effort, which the results achieved with only one hidden layer or even

without hidden layers show to be useless.

4. Discussion

From the coordinates of the target-output pairs obtained with the best-performing model (umberto-commoncrawl-cased-v1, "minimal" task-specific architecture, batch 50, 10 epochs), I calculated the error in km using the Haversine distance. From the error histogram we see how the average error (117.21 km) is not at all generalizable to all examples. In fact, 25% of the inputs were geolocated more than 154km away from the target.

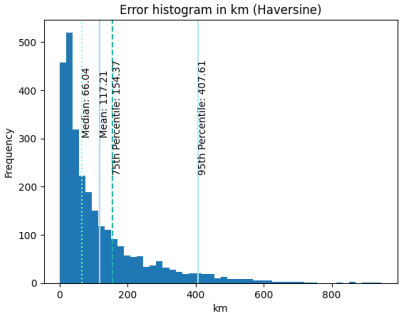


Figure 1: Error distribution of tweets; umberto-commoncrawl-cased-v1 "minimal" architecture; 50 bins

There are no specific areas where the inputs have a larger error. In contrast, inputs from areas in Piedmont-Lombardy-Veneto and Lazio-Campania have lower error than the others. In fact, two or three marked clusters can be observed in the scatterplots of the outputs (Figure 2), depending on the model configuration, the most persistent of which is between Lazio and Campania, then a cluster that follows the Alpine arc and finally less frequently by a cluster on Sicily. Excluding that this can be attributed to an imbalance in our fine-tuning dataset, this result must come from the representation of the embeddings of each model. Ramponi and Casula [1] argue on the fact that the pre-training material that had been used by those models may include content in language varieties of Italy, and they attribute it to the over-prediction of Italian of current language identifiers, observing that content both partially and fully written in language varieties of Italy is typically classified as standard italian by the Twitter language identifier. I can further hypothesize that the varieties from the areas with the smallest error are also quantitatively more present in the pre-training dataset of each model, as these are also the ones from the most densely populated areas in Italy.

However, it is very complex to reconnect these observations to one or more linguistic facts concerning the Italian regional varieties. The question then is how did we get this results, or even *"Does BERT make any sense?"* [14]. We defer the answer to later work.

Comparing the outputs' scatterplots (Figure 2) of the models umberto-commoncrawl-cased-v1, bert-base-italian-cased and AlBERTo-it fine-tuned on the provided dataset, we hope to probe in some way the embedding space generated by each model, and hopefully to gain insights into the quality and characteristics of the learned embeddings.

In the "minimal" architecture, AlBERTo-it shows a main cluster in the middle, from which all the other outputs radiate, while the other two models identify about two or three main clusters, like we said before, around the Alpine arc, between Lazio and Campania and the last, least marked, approximately on Sicily. Comparing the specifications of each model, we see that the most important differences are in the training dataset. While umberto-commoncrawl-cased-v1[1] and bert-base-italian-cased were pre-trained on a miscellaneous corpus, respectively OSCAR [15] Italian subcorpus and Wikipedia and OPUS corpus [16], AlBERTo-it was pre-trained on tweets. In this work, the models pre-trained on a miscellaneous corpus (umberto-commoncrawl-cased-v1 and bert-base-italian-cased) provided embeddings that performed better on tweets than a model pre-trained on a corpus specifically of the same genre. It is also worth noting that in Ramponi and Casula's work [1], the best performing model was AlBERTo-it, that has a vocabulary size of 128k, while in the current work the best performing ones are the 32k umberto-commoncrawl-cased-v1 and the 31k bert-base-italian-cased.

Adding a single large hidden layer makes fewer outliers compared to the models with the minimal architecture, and seems to be able to overcome at least partially the cluster of the Alpine arc and Sicily, while keeping the central one very compact. However it is necessary to remember that only umberto-commoncrawl-cased-v1 has been tested with a large hidden layer, and further experiments could be carried out.

5. Conclusions

The behavior shown by our models (need for regularization in the presence of numerous layers, better results with a single bigger hidden layer) is what we can expect from a simple neural network. However, it is astonishing that such a simple architecture manages to obtain non-disastrous results in an complex NLP task. The success of this regression task is undoubtedly attributable

[1]The OSCAR [15] subcorpus also has some subsets in other italian language varieties (such Piedmontese), but the official umberto-commoncrawl-cased-v1 model card says that it was pre-trained only on the Italian subcorpus, deduplicated.

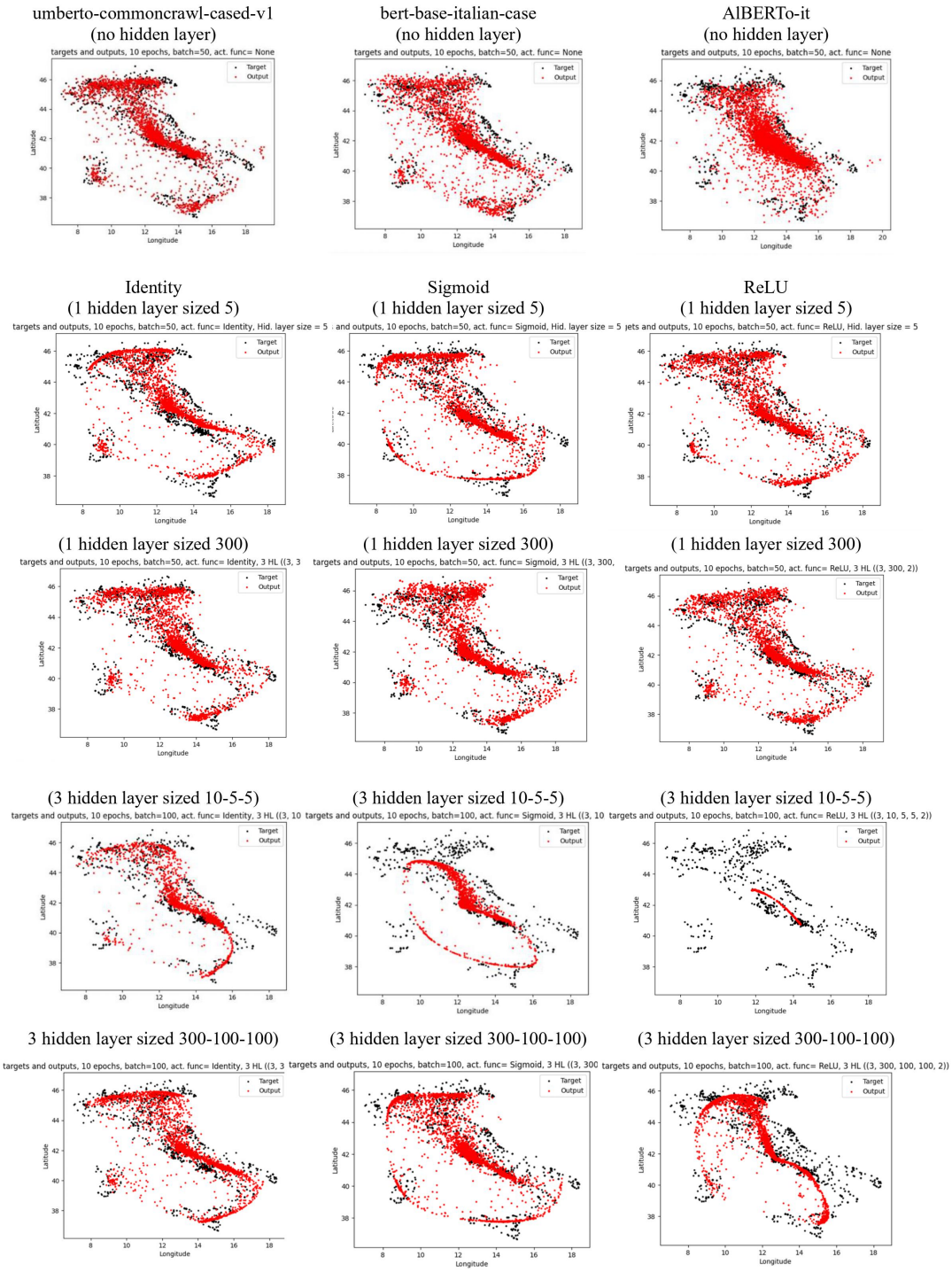


Figure 2: Scatterplots of the target (in black) and output (in red) coordinates for each configuration

to the high-level representations of the input data, together with BERT's ability to understand the linguistic context. Therefore, less is more: a simple setup, using even just two output neurons, seems to work better than a more complex one for BERT fine-tuned models on this task. Furthermore, in this work the models pre-trained on a miscellaneous corpus provides embeddings that performed better on tweets than a corpus specifically of the same genre. In conclusion, it is difficult to say how close we came to the goal, if the goal was to adequately map the diatopic variation of contemporary Italian, trying to automatically extract regional and dialectal patterns. Even if in this work we were unable to further probe the linguistic information used to carry out our task, the studies converge in believing that *BERT's structure is, however, linguistically founded, although perhaps in a way that is more nuanced than can be explained by layers alone* [17].

References

[1] A. Ramponi, C. Casula, DiatopIt: A corpus of social media posts for the study of diatopic language variation in Italy, in: Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023), Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 187–199. URL: https://aclanthology.org/2023.vardial-1.19.

[2] K. Lutsai, C. H. Lampert, Geolocation predicting of tweets using bert-based models, 2023. arXiv:2303.07865.

[3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[4] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[5] A. Ramponi, C. Casula, GeoLingIt at EVALITA 2023: Overview of the geolocation of linguistic variation in Italy task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[6] B. Han, A. Rahimi, L. Derczynski, T. Baldwin, Twitter geolocation prediction shared task of the 2016 workshop on noisy user-generated text, in: Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT), The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 213–217. URL: https://aclanthology.org/W16-3928.

[7] M. Gaman, D. Hovy, R. T. Ionescu, H. Jauhiainen, T. Jauhiainen, K. Lindén, N. Ljubešić, N. Partanen, C. Purschke, Y. Scherrer, M. Zampieri, A report on the VarDial evaluation campaign 2020, in: Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects, International Committee on Computational Linguistics (ICCL), Barcelona, Spain (Online), 2020, pp. 1–14. URL: https://aclanthology.org/2020.vardial-1.1.

[8] B. R. Chakravarthi, G. Mihaela, R. T. Ionescu, H. Jauhiainen, T. Jauhiainen, K. Lindén, N. Ljubešić, N. Partanen, R. Priyadharshini, C. Purschke, E. Rajagopal, Y. Scherrer, M. Zampieri, Findings of the VarDial evaluation campaign 2021, in: Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects, Association for Computational Linguistics, Kiyv, Ukraine, 2021, pp. 1–11. URL: https://aclanthology.org/2021.vardial-1.1.

[9] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, S. Bowman, GLUE: A multi-task benchmark and analysis platform for natural language understanding, in: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 353–355. URL: https://aclanthology.org/W18-5446. doi:10.18653/v1/W18-5446.

[10] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. arXiv:1907.11692.

[11] L. Parisi, S. Francia, P. Magnani, Umberto: an italian language model trained with whole word masking, https://github.com/musixmatchresearch/umberto, 2020.

[12] S. Schweter, Italian bert and electra models, 2020. URL: https://doi.org/10.5281/zenodo.4263142. doi:10.5281/zenodo.4263142.

[13] M. Polignano, P. Basile, M. de Gemmis, G. Semeraro, V. Basile, AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets, in: Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019), volume 2481, CEUR, 2019. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85074851349&partnerID=40&md5=7abed946e06f76b3825ae5e294ffac14.

[14] G. Wiedemann, S. Remus, A. Chawla, C. Biemann, Does BERT make any sense? interpretable word sense disambiguation with contextualized embeddings, CoRR abs/1909.10430 (2019). URL: http://arxiv.org/abs/1909.10430. arXiv:1909.10430.

[15] J. Abadji, P. Ortiz Suarez, L. Romary, B. Sagot, Towards a Cleaner Document-Oriented Multilingual Crawled Corpus, arXiv e-prints (2022) arXiv:2201.06642. arXiv:2201.06642.

[16] J. Tiedemann, S. Thottingal, OPUS-MT – building open translation services for the world, in: Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, European Association for Machine Translation, Lisboa, Portugal, 2020, pp. 479–480. URL: https://aclanthology.org/2020.eamt-1.61.

[17] J. Niu, W. Lu, G. Penn, Does BERT rediscover a classical NLP pipeline?, in: Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 3143–3153. URL: https://aclanthology.org/2022.coling-1.278.

LangLearn at EVALITA 2023: Overview of the Language Learning Development Task

Chiara Alzetta[1], Dominique Brunato[1], Felice Dell'Orletta[1], Alessio Miaschi[1], Kenji Sagae[2], Claudia H. Sánchez-Gutiérrez[2] and Giulia Venturi[1]

[1]ItaliaNLP Lab, CNR, Istituto di Linguistica Computazionale 'A.Zampolli', Pisa, Italy
[2]University of California, Davis

Abstract
Language Learning Development (LangLearn) is the EVALITA 2023 shared task on automatic language development assessment, which consists in predicting the evolution of the written language abilities of learners across time. LangLearn is conceived to be multilingual, relying on written productions of Italian and Spanish learners, and representative of L1 and L2 learning scenarios. A total of 9 systems were submitted by 5 teams. The results highlight the open challenges of automatic language development assessment.

Keywords
Language learning development, student essays, shared task, multilingual language learning assessment

1. Introduction

Over the last twenty years, there has been a growing interest in exploiting the potential of Natural Language Processing (NLP) tools to characterize the properties of learners' language and to study how it evolves over time, both in first (L1) and second language (L2) acquisition scenarios. A similar concern has been paid to turning theoretical considerations into educational applications, such as Intelligent Computer-Assisted Language Learning (ICALL) systems [1] and tools for automatically scoring learners' writing with respect to language proficiency and writing quality [2, 3], and more generally systems able to automatically assign a learner's language production to a given developmental level [4, 5] or to operationalize sophisticated metrics of language development thus alleviating the laborious manual computation of these metrics by experts [6, 7, 8].

Generally, a greater number of studies has been carried out in the field of L2 learning where the study of L2 writings is seen as a proxy for language ability development [9]. In this respect, much work is devoted to predicting the degree of L2 proficiency according to expert-based evaluation [10] or to modeling the evolution of grammatical structures' competence with respect to predefined grades, e.g. the Common European Framework of Reference for Languages (CEFRL) [11, 12, 13]. On the other side, fewer studies have focused on exploiting NLP techniques in the context of L1 development. Among these, we can mention studies devoted to assessing syntactic development in preschool children [8, 14] and to examine overall writing ability and its development during later language acquisition [15].

It is worth noting that most research has been focused on English. Few exceptions are represented by e.g. the works by [16] and [17], which investigated writing development in German-speaking students across the elementary and secondary school, and [18], and [19], who proposed a methodology for tracking the evolution of written language competence of L2 Spanish and L1 Italian learners, respectively.

The Language Learning Development Task (LangLearn) organized at EVALITA 2023 [20] continues this scenario and it represents the first shared task on automatic language development assessment. It was aimed at developing and evaluating systems to predict the evolution of the written language abilities of learners across several time intervals. Additionally, the task was conceived to be multilingual, relying on written productions of Italian and Spanish learners, and representative of L1 and L2 learning scenarios.

2. Definition of the task

The assessment of language development is cast in LangLearn as a binary classification problem: it consists in predicting the relative order of two essays written by the same student. We started from the assumption that, given a set of chronologically ordered essays written by the same student, a document d_j should have a higher

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
chiara.alzetta@ilc.cnr.it (C. Alzetta);
dominique.brunato@ilc.cnr.it (D. Brunato);
felice.dellorletta@ilc.cnr.it (F. Dell'Orletta);
alessio.miaschi@ilc.cnr.it (A. Miaschi); sagae@ucdavis.edu (K. Sagae); chsanchez@ucdavis.edu (C. H. Sánchez-Gutiérrez);
giulia.venturi@ilc.cnr.it (G. Venturi)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper18.pdf

quality level with respect to the ones written previously (d_i). Specifically, we followed the approach devised by [19]: given a randomly ordered pair of essays (e_1, e_2) written by the same student, we ask to predict whether e_2 was written before e_1.

LangLearn was articulated in two sub-tasks based on the resources allowed for training the models.

- **Sub-task 1** consists in predicting the order of essays using only the official training data released for the task;
- **Sub-task 2** consists in predicting the order of essays using information acquired from the training data released for the task and also from additional external resources.

3. Datasets

In line with the aim of having a multilingual shared task, we distributed two datasets composed of essays written by learners of the Italian and Spanish languages. Notably, the two datasets reflect an additional dimension of variation, which is the different learning scenarios from which the written productions were obtained. Specifically, the collection of Italian essays was written by students learning Italian as their first language, while the Spanish essays were produced by L2 learners.

For each corpus, LangLearn participants were provided with two files:

- a .tsv file containing the following information pertaining to a pair of essays (e_1, e_2) written by the same student: IDs of e_1 and e_2 in the correct chronological order, and t_1 and t_2 corresponding to the time of writing of e_1 and e_2.
- an XML file containing the text of the essays with randomly generated document IDs, as in the example below:

```
<dataset>
<doc id="9843"> Essay </doc>
<doc id="7432"> Essay </doc>
</dataset>
```

3.1. Corpus Italiano di Apprendenti L1 (*CItA*)

CItA (*Corpus Italiano di Apprendenti L1*) [21] is a longitudinal corpus of essays written by the same L1 Italian students in the first (2012-2013) and second year (2013-2014) of lower secondary school. The original corpus contains a total of 1,352 essays written by 156 students. The essays belong to five textual typologies, which reflect the different prompts students were asked to respond to, i.e. reflexive, narrative, descriptive, expository and argumentative.

For the purposes of the LangLearn shared task, we selected a subset of 882 essays authored by 133 different students at different time intervals. A time interval is identified by the year and specific period during which each essay was produced (e.g., the label 1_4 denotes the fourth essay written during the first year). Specifically, we considered 11 intervals, six for the first year and five for the second one. As it can be seen in Figure 1, the essays feature diverse linguistic characteristics across the considered time intervals[1]. In fact, essays written in the first year tend to be shorter in terms of the total number of tokens than those produced in the second year. Interestingly, the length of the document is a raw text feature highly related to various linguistic aspects that shape the writing style of an essay. Furthermore, the essays are increasingly lexically richer across time, as emerged from the Type/token ratio (TTR) values calculated for the first 100 tokens of the texts. It is worth noting that the last essays of the second year (interval 11) deviate from this trend. This is possibly due to the fact that they are mostly related to similar prompts that involved completing a history. In this case, students tend to write shorter and less lexically varied essays.

In order to build the training and test sets of LangLearn, essays from each student were paired based on their chronological order of writing, ensuring that the first essay in each pair was written prior to the second. This process resulted in 2,673 essay pairs: 2,366 were assigned to the train set, and the remaining 307 were placed in the test set. The distribution of pairs across time intervals is reported in Table 1. Note that some time interval pairs (e.g. $1_2 - 1_5$) appear only in the test set. This is done to challenge participants since they do not have any corresponding pairs within the train set. Similarly, we isolated 4 students whose essays appear only in the test set, while the essays of 49 students appear only in the train set and the essays of 80 students appear in both sets. Indeed, it is possible for the same essay to appear in both the training and test sets, but it would appear in different pairs, ensuring that a specific pair occurs exclusively in either the train or test set.

3.2. Corpus of Written Spanish of L2 and Heritage Speakers (*COWS-L2H*)

The COWS-L2H (*Corpus of Written Spanish of L2 and Heritage Speakers*) corpus [22] consists of 3,498 short essays written by second language (L2) students enrolled in one of ten lower-division Spanish courses at a single American university. Student compositions in the corpus are

[1]Note that these two linguistic characteristics are those used to compute the baseline scores.

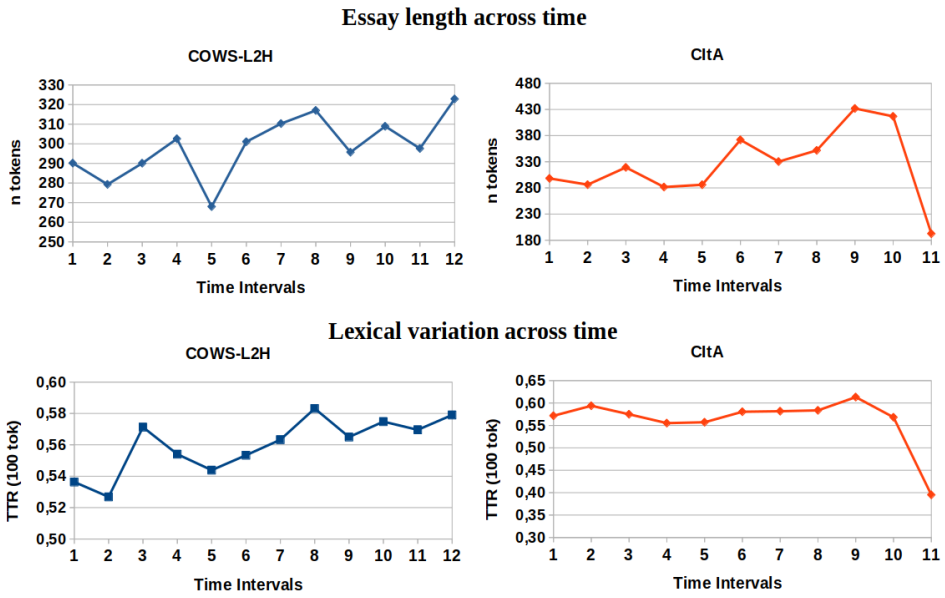


Figure 1: Distribution of two linguistic characteristics of the LangLearn datasets across the considered time intervals.

written in response to one of six writing prompts, which are changed periodically. According to these prompts, in each essay, the student is asked to write about: a famous person, your perfect vacation plan, a special person in your life, a terrible story, a description of yourself, or a beautiful story. During each period (an academic quarter, which consists of ten weeks of courses) of data collection, students are asked to submit two compositions, approximately one month apart, in response to the previously mentioned prompts. These composition themes are designed to be relatively broad, to allow for a wide degree of creative liberty and open-ended interpretation by the writer.

To select essays from the original COWS-L2H dataset for the LangLearn task, we considered only essays written by students who wrote essays in two separate academic terms. This way, we can pair essays written at different points in time by the same student. To reduce the possibility that factors independent of language learning could systematically differentiate between essays in a pair, we considered only pairs of essays written in response to the same prompt. With these constraints, we were left with 1,329 pairs of essays written by 440 students. To split these essay pairs into training and test sets, we selected the essays written by 330 students to be in the training set, and the essays written by the remaining 110 students to be in the test set. This means that, in contrast with the CItA dataset used in LangLearn, there is no overlap in essays or authors between the training and test sets. The resulting training set contains 1,009 essay pairs, and the test set contains 320 essay pairs. The time interval between essays in a pair usually consists of one, two or three academic terms, with each term corresponding to 10 weeks of courses (Table 2). It is important to note that these intervals are not easily comparable across datasets, since COWS-L2H deals with highly structured L2 instruction, which progresses differently from L1 writing.

4. Evaluation

Baseline The baseline scores were calculated by training a LinearSVM using, for each pair (e_1, e_2), the number of tokens per document (in each pair) and the type/token ratio of the first 100 tokens in each document as input features.

Metrics The models' performance achieved on the CItA and COW-L2H test sets have been evaluated independently using Accuracy (A) and F1-score (F-score).

5. Submitted Systems and Participants

Following a call for interest, 5 teams registered for the task and submitted their predictions for both datasets, for a total of 18 runs (namely, 9 for each language tackled in the shared task). Eventually, one team (i.e. aroyehun_angel) did not submit a system report, thus we included

CItA - Train set											
	1_1	1_2	1_3	1_4	1_5	1_6	2_1	2_2	2_3	2_4	2_5
1_1		86	35	50	53	11	85	72	86	80	31
1_2			34	50			87		87	79	31
1_3				36	36	15	37	37	36	45	
1_4					56	16	57	53	55	50	
1_5						18	62	58	60	52	
1_6							22	15	17	19	
2_1								71	95	84	30
2_2									76	65	
2_3										80	28
2_4											28

CItA - Test set											
	1_1	1_2	1_3	1_4	1_5	1_6	2_1	2_2	2_3	2_4	2_5
1_1		2					2	2	2	1	6
1_2			2	1	54	15	4	74	4	4	6
1_3				1	2	2	2	4	2	3	
1_4					1	1	1	2	1	1	
1_5						2	2	4	2	3	
1_6							2	4	2	3	
2_1								6	4	4	6
2_2									6	7	35
2_3										4	6
2_4											3

Table 1
Distribution of essay pairs with respect to time intervals in the train and test sets of CItA. The rows correspond to the time points of the first essay in a pair, and the columns the time points of the second essay in a pair.

their scores in the overall dashboard, but we excluded them from the system description and error analyses. As shown in Table 3, all teams participated only in sub-task 1.

BERT_4EVER [23] proposed three different systems based on the base Italian BERT[2] model [24]. For fine-tuning the models, the team augmented the CItA and COWS-L2H datasets by reversing essay pairs to obtain negative examples and generating new positive examples by constructing transitive pairs. In the first system, BERT, BERT was fine-tuned performing simultaneous training on the augmented CItA and COWS-L2H datasets. The second model, Sequential, employs a novel sequential information attention mechanism to capture the interaction between the essays in a pair, which allows for incorporating the attention weights derived from the last-written essay in the representation of the pair relying on the [CLS] token and using average pooling. This pair representation is then fed into a linear classifier with a softmax function. The third model proposed is the Merge one, which fuses BERT and Sequential by averaging their output probabilities.

[2]https://huggingface.co/dbmdz/bert-base-italian-uncased

COWS-L2H - Train set										
	SU17	F17	W18	S18	W19	S19	F19	S20	F20	W21
S17	2	45	24	12						
SU17		2								
F17			51	25						
W18				74						
S18										
SU18										
F18					153	65	37			
W19						109	45			
S19							61			
SU19										
F19										
W20								109	24	36
S20									35	35
SU20										
F20										65

COWS-L2H - Test set										
	SU17	F17	W18	S18	W19	S19	F19	S20	F20	W21
S17		17	13	5						
SU17										
F17			17	11						
W18				24						
S18										
SU18										
F18					43	19	11			
W19						45	18			
S19							21			
SU19										
F19										
W20								29	6	6
S20									9	11
SU20										
F20										15

Table 2
Distribution of essay pairs with respect to time intervals in the train and test sets of COWS-L2H. The rows correspond to the time points of the first essay in a pair, and the columns the time points of the second essay in a pair.

bot.zen [25] tackled LangLearn as a regression problem, where the goal was to determine the stage of the learning process at which a student wrote a text. To achieve this, the team first pre-processed the official training sets in order to acquire the absolute order of each essay written by a student. Then, they performed predictions relying on an ensemble of decision tree algorithms. The model was trained using 125 normalised features capturing lexical and morpho-syntactic properties for each essay. By using MALT-IT2 [26], the team was able to include a set of features measuring text complexity in terms of document length, and lexical, syntactic, and morpho-syntactic properties. These features, however, are available only for the Italian language, thus they were used only for CItA predictions.

Team	Members	Affiliation	Sub-task	Runs per Language
BERT_4EVER	4	Guangdong University of Foreign Studies, Guangdong University of Technology, China	1	3
aroyehun_angel	2	Instituto Politecnico Nacional, Mexico	1	2
bot.zen	6	Eurac Research, Italy	1	1
IUSSnets	3	Iuss Pavia, Italy	1	1
ExtremITA	4	Universitá degli Studi di Roma Tor Vergata, Universitá di Torino, Italy	1	2

Table 3
Teams participating in EVALITA 2023 LangLearn shared task. For each team, we detail the number of team members, their affiliations, the sub-task(s) they participated in, and the number of submitted runs for each language of the shared task.

IUSS-Nets [27] approached LangLearn using linguistics features (e.g. density of various part-of-speech categories, frequency of different kinds of syntactic constituents, mean sentence length, etc.) extracted using the existing Common Text Analysis Platform, or CTAP [28], and surprisal-based metrics derived from token probabilities obtained using pretrained language-specific BERT models. These different pieces of information were encoded in features used in random forest classifiers. Interestingly, unlike most systems in LangLearn, which obtained better performance on the CItA dataset than on COWS-L2H, this approach produced higher accuracy and F-Score on COWS-L2H. In fact, it produced the strongest results on the COWS-L2H dataset among those submitted. Although its performance on CItA was not among the strongest submitted, it was still substantially above the baseline.

ExtremITA [29] team participated in the task with two Language Models trained in a multi-task learning framework. The first model is an encoder-decoder based on IT5-small [30], while the second model was a decoder based on Camoscio [31], the Italian version of LLaMA [32]. These models show substantial differences in terms of parameter count, with IT5-small comprising around 110 million parameters, whereas the utilized version of Camoscio encompasses 7 billion parameters. Both models underwent joint fine-tuning on all EVALITA 2023 tasks and sub-tasks, leveraging prompting techniques. Specifically, for the LangLearn task, the extremIT5 model received each instance of the dataset with the task name preceding it as input, and it produced the predicted label as output. Conversely, the extremITLLaMa model, which requires a structured prompt, was provided with a textual description of the task and the desired output format specification, as follows: *"Questi due testi separati da [SEP] sono presentati nell'ordine in cui sono scritti? Rispondi sì o no"*. As regards the dataset treatment, some preprocessing steps were adopted: firstly, the dataset was segmented into sentences, allowing a maximum of 100 tokens per sentence. Additionally, in order to augment the dataset, inverted sentence pairs were incorporated, resulting in an expansion of the dataset from 3,377 to 6,438 examples.

CItA		
Team	**Accuracy**	**F-Score**
BERT_4EVER-BERT	0.932	0.934
BERT_4EVER-Merge	0.925	0.927
BERT_4EVER-Sequential	0.925	0.926
aroyehun_angel-system2	0.863	0.865
aroyehun_angel-system1	0.840	0.845
bot.zen	0.834	0.837
IUSS-Nets	0.645	0.673
ExtremITA-camoscio-lora	0.596	0.613
Baseline	0.550	0.549
ExtremITA-it5	0.606	0.410
COWS-L2H		
Team	**Accuracy**	**F-Score**
IUSS-Nets	0.753	0.752
aroyehun_angel-system1	0.703	0.708
BERT_4EVER-Sequential	0.641	0.663
Baseline	0.659	0.663
BERT_4EVER-Merge	0.616	0.631
BERT_4EVER-BERT	0.609	0.620
aroyehun_angel-system2	0.588	0.569
ExtremITA-camoscio-lora	0.575	0.553
bot.zen	0.497	0.517
ExtremITA-it5	0.506	0.160

Table 4
LangLearn shared task leaderboard.

6. Results

Table 4 reports the leaderboard of systems participating in the LangLearn shared task. Most systems outperformed the baseline when tested on CItA dataset while surpassing the baseline proved to be more challenging on COWS-L2H dataset. The team BERT_4EVER submitted the best-performing systems in the L1 scenario, while the highest score for the Spanish dataset was achieved by the IUSS-Nets team. ExtremITA obtained the lowest

scores on both datasets.

Overall, we observe varying system rankings across the two learning scenarios. We discuss such variation in more depth in the next Section.

7. Discussion

Upon examination of system performance, we notice differences in model performance between the CItA and COWS-L2H datasets. Considering that each dataset reflects a different learning scenario, this might indicate that the challenges posed by these scenarios were distinct. One notable finding is that models leveraging stylistic properties of essays, such as the IUSS-Nets model, were more effective in the L2 setting. On the other hand, teams that employed Neural Language Models achieved higher results on the CItA dataset.

The observed differences in performance might be attributed to two main factors: model architectures and specific properties of the two learning scenarios. Concerning the former, we highlight, for instance, that the BERT model used by the BERT_4EVER team was pretrained only on Italian texts. This choice likely contributed to its lower performance on COWS-L2H, despite the simultaneous fine-tuning on both CItA and COWS-L2H. In fact, while BERT was the best-performing model of the BERT_4EVER team and overall on CItA, it was surpassed on Spanish essays by their Sequential model, which incorporates information about the interaction between the essays in a pair. Similar observations can be made for the bot.zen and IUSS-Nets teams. Both teams employed classification models that leverage a set of explicit features capturing linguistic properties of the texts. While both teams exploited features measuring raw text properties and the distribution of part-of-speech and syntactic dependencies for both languages, they differed in terms of features that captured deeper textual properties. Specifically, IUSS-Nets achieved the highest score on the COWS-L2H dataset thanks to a wide set of features measuring text complexity, sophistication, refinement, lexical variety, and cohesion. Conversely, the bot.zen team was unable to compute features capturing text complexity for Spanish, resulting in lower scores for that language.

These results reflect also specific properties of the two learning scenarios of LangLearn, which clearly affected all systems submitted to the shared task. As observed by [27], the evolution of writing abilities in a second language shows greater variation in terms of style within a shorter time period compared to a first language. We can assume that during the learning phase of an L2, new linguistic structures are acquired by the students in a highly structured schedule dictated by the L2 learning environment, gradually becoming more complex in a somewhat uniform way. Consequently, the essays produced by L2 learners may primarily serve as a measure of their progress in acquiring these new, more complex structures. On the other hand, L1 learners may face challenges from their teachers to enhance their proficiency in accurately using linguistic structures they have already acquired. As a consequence, L2 essays may exhibit more significant stylistic variations as learners are faced with the acquisition of new language structures. In contrast, L1 essays over time may show a more accurate use of already familiar linguistic structures, highlighting the learners' mastery of these elements.

Team	Accuracy	F-Score
BERT_4EVER-BERT	0.855	0.865
BERT_4EVER-Merge	0.841	0.850
BERT_4EVER-Sequential	0.841	0.846
bot.zen	0.717	0.717
IUSS-Nets	0.648	0.683
ExtremITA-camoscio-lora	0.600	0.608
Baseline	0.566	0.559
ExtremITA-it5	0.655	0.537

Table 5
Results on the CItA Test set considering only unseen students.

To deepen our analyses on the CItA dataset, we compared the system performance on a subset of essay pairs that correspond to the most challenging prediction scenario, i.e. considering pairs involving students whose essays appear only in the test set. The results on this subset are reported in Table 5. As can be noted, the system ranking remains unvaried, but the bot.zen and BERT_4EVER systems suffer a drop in their performance on this setting. The main cause of the decline in scores is due to the increased complexity of this particular setting. In fact, systems cannot rely on information extracted from essays present both in the training and test sets, although paired with different essays. As a result, the systems must rely solely on their generalization abilities to discern significant variations within each essay pair. However, it is important to acknowledge that even in this particular setting, the scores achieved by the BERT_4EVER team significantly surpass the baseline. This further highlights the potential of language models, particularly in the L1 classification scenario, as previously mentioned.

As a final remark, it is worth discussing the performance of ExtremITA systems. This team employed two Large Language Models to tackle all shared tasks proposed in the EVALITA 2023 campaign and explored the applicability of a single model in solving multiple different tasks. Although extremITLLaMA achieved the top position in 41% of all EVALITA sub-tasks (i.e., 13 out of 22 sub-tasks) and a top-three placement in 14 sub-tasks, the results on LangLearn were just slightly above the baseline on CItA and below the baseline on COWS-L2H. Such a result lays the foundation for an interesting and highly

timely discussion on the effectiveness of these large and powerful models on real-world tasks. It appears, in fact, that tasks that are strongly affected by stylistic properties, such as language learning development assessment, still pose challenges to these models.

8. Conclusions

In this report, we introduced LangLearn, the first shared task dedicated to the development of systems able to automatically predict the development of language learning starting from learners' essays, in two learning scenarios and in a multilingual setting. Analysis of the results from the 9 submitted models indicates that the task of language learning development assessment continues to present numerous unresolved challenges. Notably, models that relied on explicit stylistic features demonstrated superior performance in Spanish as an L2 learning scenario. Conversely, Large Language Models showcased greater effectiveness in Italian as an L1 learning scenario.

These findings shed light on the complex nature of language learning assessment and suggest possible future directions for future evaluation campaigns. On the one hand, by leveraging insights from the LangLearn task, researchers can devise new approaches that incorporate both explicit stylistic features and the strengths of Large Language Models. On the other hand, the comparably lower scores achieved by the ExtremITA in our task seem to prompt a new typology of evaluation campaigns devoted to putting under pressure the potential of Large Language Models, pushing the boundaries of their language comprehension and generation capabilities.

Acknowledgments

The authors gratefully acknowledge the support of the PNRR MUR project PE0000013-FAIR.

References

[1] S. Granger, Error-tagged learner corpora and call: A promising synergy, CALICO journal (2003) 465–480.

[2] D. S. McNamara, S. A. Crossley, R. D. Roscoe, L. K. Allen, J. Dai, A hierarchical classification approach to automated essay scoring, Assessing Writing 23 (2015) 35–59.

[3] P. Deane, T. Quinlan, What automated analyses of corpora can tell us about students' writing skills, Journal of Writing Research 2 (2010) 151–177.

[4] K. Sagae, A. Lavie, B. MacWhinney, Automatic measurement of syntactic development in child language, in: Proceedings of the ACL, ACL, 2005, pp. 197–204.

[5] X. Lu, Automatic measurement of syntactic complexity in child language acquisition, International Journal of Corpus Linguistics 14 (2009) 3–28.

[6] B. Bram, A. Housen, Conceptualizing and measuring short-term changes in l2 writing complexity, Journal of Second Language Writing 26 (2014) 42–65.

[7] S. A. Crossley, D. McNamara, Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners, Journal of Second Language Writing 26 (2014) 66–79.

[8] S. Lubetich, K. Sagae, Data-driven measurement of child language development with simple syntactic templates, in: Proceedings of COLING: Technical Papers, 2014, pp. 2151–2160.

[9] S. Crossley, Linguistic features in writing quality and development: An overview, Journal of Writing Research (2020).

[10] S. A. Crossley, D. S. McNamara, Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication, Journal of Research in Reading 35 (2012) 115–135.

[11] S. Vajjala, K. Loo, Automatic CEFR level prediction for Estonian learner text, in: Proceedings of the third workshop on NLP for computer-assisted language learning, 2014, pp. 113–127.

[12] E. Volodina, I. Pilán, D. Alfter, et al., Classification of Swedish learner essays by CEFR levels, in: CALL communities and culture–short papers from EUROCALL, 2016, pp. 456–461.

[13] L. Zilio, R. Wilkens, C. Fairon, An SLA corpus annotated with pedagogically relevant grammatical structures, in: Proceedings of LREC, European Language Resources Association (ELRA), Miyazaki, Japan, 2018.

[14] K. Sagae, Tracking child language development with neural network language models, Frontiers in Psychology 12 (2021).

[15] A. S. Crossley, J. Weston, S. McLain Sullivan, D. S. McNamara, The development of writing proficiency as a function of grade level: A linguistic analysis., Written Communication, Written Communication, vol. 28, no. 3, pp. 282–311 (2011).

[16] Z. Weiss, D. Meurers, Analyzing linguistic complexity and accuracy in academic language development of german across elementary and secondary school, in: Proceedings of BEA, 2019, pp. 380–393.

[17] E. Kerz, Y. Qiao, D. Wiechmann, M. Ströbel, Becoming linguistically mature: Modeling English and German children's writing development across school grades, in: Proceedings of BEA, ACL, Online, 2020, pp. 65–74.

[18] A. Miaschi, S. Davidson, D. Brunato, F. Dell'Orletta, K. Sagae, C. H. Sanchez-Gutierrez, G. Venturi, Tracking the evolution of written language competence in L2 Spanish learners, in: Proceedings of BEA, ACL, Online, 2020, pp. 92–101.

[19] A. Miaschi, D. Brunato, F. Dell'Orletta, A nlp-based stylometric approach for tracking the evolution of l1 written language competence, Journal of Writing Research 13 (2021) 71–105.

[20] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[21] A. Barbagli, Quanto e come si impara a scrivere nel corso del primo biennio della scuola secondaria di primo grado, Nuova Cultura, 2016.

[22] S. Davidson, A. Yamada, P. Fernandez Mira, A. Carando, C. H. Sanchez Gutierrez, K. Sagae, Developing NLP tools with a new corpus of learner spanish, in: Proceedings of the 12th LRE Conference, ELRA, Marseille, France, 2020, pp. 7240–7245.

[23] H. Wu, N. Lin, S. Jiang, L. Xiao, BERT_4EVER at LangLearn: Language development assessment model based on sequential information attention mechanism, in: M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi (Eds.), Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, September 7th-8th 2023, Parma, 2023.

[24] T. Wolf, L. Debut, V. Sanh, alii, Transformers: State-of-the-art natural language processing, in: Proc. of EMNLP, ACL, Online, 2020, pp. 38–45.

[25] E. W. Stemle, M. Tebaldini, F. Bonanni, F. Pellegrino, P. Brasolin, G. H. Franzini, J.-C. Frey, O. Lopopolo, S. Spina, bot.zen at LangLearn: regressing towards interpretability, in: M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi (Eds.), Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, September 7th-8th 2023, Parma, 2023.

[26] V. Santucci, F. Santarelli, L. Forti, S. Spina, Automatic classification of text complexity, Applied Sciences 10 (2020) 7285.

[27] M. Barbini, E. Zanoli, C. Chesi, IUSS-Nets at LangLearn: The role of morphosyntactic features in language development assessment, in: M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi (Eds.), Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, September 7th-8th 2023, Parma, 2023.

[28] X. Chen, D. Meurers, CTAP: A web-based tool supporting automatic complexity analysis, in: Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC), The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 113–119.

[29] C. D. Hromei, D. Croce, V. Basile, R. Basili, ExtremITA at EVALITA 2023: Multi-task sustainable scaling to large language models at its extreme, in: M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi (Eds.), Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, September 7th-8th 2023, Parma, 2023.

[30] G. Sarti, M. Nissim, IT5: Large-scale text-to-text pretraining for italian language understanding and generation, ArXiv preprint 2203.03759 (2022). URL: https://arxiv.org/abs/2203.03759.

[31] A. Santilli, Camoscio: An italian instruction-tuned llama, https://github.com/teelinsan/camoscio, 2023.

[32] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. arXiv:2302.13971.

BERT_4EVER at LangLearn: Language Development Assessment Model based on Sequential Information Attention Mechanism

Hongyan Wu[1], Nankai Lin[2,*], Shengyi Jiang[1,*] and Lixian Xiao[3]

[1]School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou, Guangdong, PR China
[2]School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, Guangdong, PR China
[3]Faculty of Asian Languages and Cultures, Guangdong University of Foreign Studies, Guangzhou, Guangdong, PR China

Abstract

In recent years, investigations into language acquisition have greatly benefited from the utilization of natural language processing technologies, particularly in analyzing extensive corpora consisting of authentic texts produced by learners across the realms of first and second language acquisition. A crucial task in this domain involves the assessment of language learners' language ability development. The "Language Learning Development" task featured in EVALITA 2023 [1] marks a significant milestone as the inaugural shared task focused on automated language development assessment, which entails predicting the relative order of two essays written by the same student. We introduce a novel attention mechanism, namely sequential information attention mechanism, with the primary objective of exploiting information interaction between sequence texts. Experimental results on the COWS dataset show the effectiveness of our proposed sequential information attention mechanism, showcasing its substantial impact on model performance during the final evaluation phase.

Keywords

Language Development Assessment, Sequential Information Attention Mechanism, BERT,

1. Introduction

Recently, there has been a surge of interest in harnessing the potential of natural language processing (NLP) tools and machine learning techniques to explore the realm of language development, both in first (L1) and second language (L2) acquisition scenarios. The primary objective revolves around comprehensively characterizing the linguistic attributes of learners and the dynamic evolution of their language ability across different modalities and stages of acquisition. The utilization of learner corpora and the enhanced dependability of linguistic features extracted through computational tools and machine learning techniques have significantly advanced our comprehension of the linguistic properties exhibited by language learners. The empirical evidence has shed light on the temporal dynamics and the evolution of these language properties as learners progress in language ability [2].

A significant focus of scholarly inquiry has been directed towards the exploration of various avenues for advancing the field of language development research. One prominent line of investigation has concentrated on the formulation and refinement of automated methodologies capable of effectively dealing with intricate metrics associated with the multifaceted process of language development. By doing so, these methodologies aim to alleviate the arduous and time-consuming manual computation that domain experts traditionally undertake in their analyses [3, 4].

In parallel, another compelling track of research has embarked upon the more demanding endeavor of employing entirely data-driven approaches [5] within the domain of language development. This approach capitalizes on the wealth of information contained within textual data, leveraging diverse linguistic features that can be automatically extracted from such data sources. By harnessing the power of machine learning and computational techniques, researchers have sought to develop robust models capable of automatically assigning a learner's language production to a specific developmental level.

The data-driven approach holds significant promise, as it offers the potential to enhance the efficiency and accuracy of language development assessments. By training models on extensive datasets containing diverse samples of language production, researchers aim to uncover patterns and associations that can inform the classification of learners into distinct developmental levels. This approach not only holds the potential to expedite the process of evaluating language development but also offers valuable insights into the underlying mechanisms and

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
*Corresponding author.
2754976781@qq.com (H. Wu); neakail@outlook.com (N. Lin); 200511402@oamail.gdufs.edu.cn (S. Jiang); 173829137@qq.com (L. Xiao)
0000-0003-2838-8273 (N. Lin)
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper19.pdf

progression of language acquisition.

The "Language Learning Development" task featured in EVALITA 2023 is concerned with predicting the chronological sequence of essays produced by the same student over different periods. We introduce a novel attention mechanism, namely sequential information attention mechanism (SIAM), intending to exploit information interaction between sequence texts. We submitted three results in total, namely the fine-tuned BERT model (Run 2), the fine-tuned BERT model with SIAM (Run 3), and the fusion of the results of the previous two models (Run 1). Experimental results demonstrate that our proposed sequential information attention mechanism has a remarkable impact on model performance during the final evaluation phase.

2. Related Work

The existing research on the language development assessment task is mainly divided into two types, one focuses on the construction of the language ability development assessment model based on language features, and the other is concerned with the construction of a language ability development assessment model based on neural networks.

Given the inherent challenge of establishing a unique indicator of linguistic complexity within the domain of second language (L2) development, a diverse range of features spanning various linguistic levels have been employed as inputs for supervised classification systems. These systems are trained on genuine learner data pertaining to different L2 languages. Notable examples include the works of Hancke and Meurers [6] as well as Vajjala and Lõo [7], which respectively investigated L2 German and L2 Estonian. Pilán and Volodina [8] provided a comprehensive analysis of predictive features extracted from both receptive and productive texts within the context of Swedish L2 acquisition. Miaschi et al. [9] used various linguistic features automatically extracted from students' written expressions to track the evolution of written language abilities of second-language Spanish learners. Furthermore, Miaschi et al. [5] proposed a natural language processing-based style measure to track the evolution of Italian L1 learners' written language competence, which relied on capturing a range of linguistically motivated features in terms of text style. In a study conducted by Bulté and Housen [10], the objective was to determine the nature and extent of English L2 writing proficiency development among 45 adult ESL learners throughout the duration of an intensive short-term academic English language program. The investigation employed quantitative measures that specifically targeted various aspects of lexical and syntactic complexity exhibited in the learners' writing performance. Additionally, the study aimed to establish a comparison between the scores obtained from these measures and the subjective ratings provided for the overall writing quality of the learners.

Recent work on the application of neural networks to language modeling has shown that models based on certain neural architectures can capture syntactic information from utterances and sentences even without explicit syntactic goals. Sagae [11] conducted a study to determine whether a fully data-driven model of language development, utilizing a recurrent neural network encoder to encode utterances, could track changes in children's language over the course of their language development in a comparable manner to the leverage of expertly established language assessment metrics for language-specific information.

The untapped potential of neural networks in language development assessment tasks necessitates further exploration, as the application of pre-trained models in this context has not been investigated.

3. Method

3.1. Overview

Figure 1 provides a comprehensive depiction of our methodology. Initially, we concatenate the historical text and current text, and utilize the pre-trained model BERT to encode them. Subsequently, we employ the sequential information attention mechanism to capture the interaction of information within the sequence of text, thereby updating the representation of the historical text to obtain an improved global representation. Ultimately, we combine the enhanced global representation of the historical text with the original sentence representation for the final assessment of language development.

3.2. Text Representation

Aiming to effectively capture the intricate semantic information embedded within the text, we employ a non-autoregressive pre-trained model BERT [12] renowned for its remarkable performance in generating text-based semantic representations for sentence encoding. BERT possesses abundant linguistic, syntactic, and lexical knowledge, which is acquired through unsupervised training on a substantial corpus during the pre-training phase. The fundamental architecture of the model encompasses a multi-layer bidirectional Transformer encoder [13], facilitating global information processing and extraction. Given a historical text $T_a = \{w_a^1, w_a^2, w_a^3, ..., w_a^n\}$ and a current text $T_b = \{w_b^1, w_b^2, w_b^3, ..., w_b^m\}$, two special tokens $[CLS]$ and $[SEP]$ of BERT are utilized to stitch them together, forming the text input $T = \{[CLS], w_a^1, w_a^2, w_a^3, ..., w_a^n, [SEP], w_b^1, w_b^2, w_b^3, ..., w_b^m, [SEP]\}$

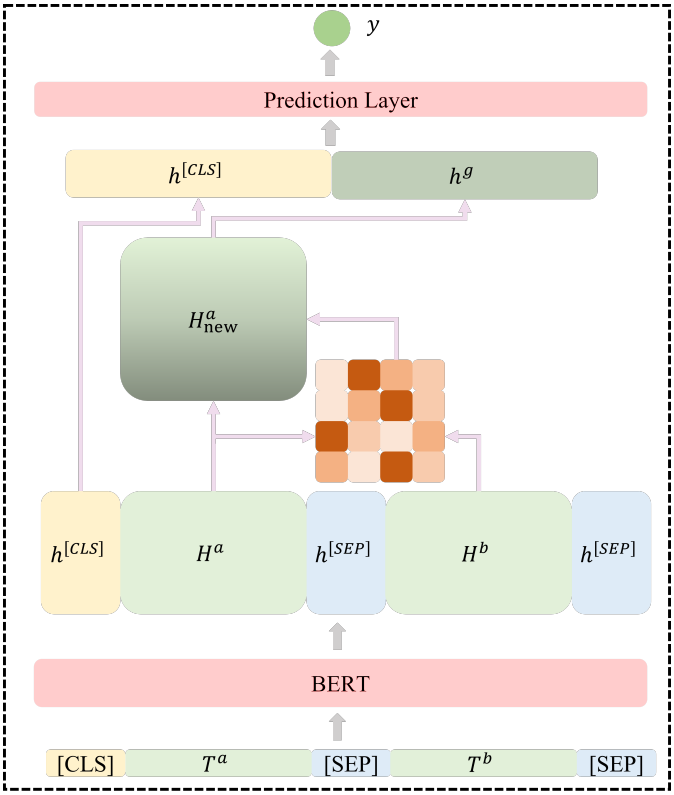


Figure 1: Model Framework Diagram.

to the pre-trained model BERT, where n and m represent the length of the two texts respectively. The semantic representation H corresponding to T encoded by the BERT pre-trained model is:

$$H = Encoder(T) \tag{1}$$

where $H = \{h^{[CLS]}, h_a^1, h_a^2, h_a^3, ..., h_a^n, h^{[SEP]}, h_b^1, h_b^2, h_b^3, ..., h_b^m, h_{[SEP]}\} \in R^{(n+m)\cdot z}$ and z denotes the dimension of semantic representation.

Then the semantic representation of text T_a and text T_b is respectively:

$$H^a = \{h_a^1, h_a^2, h_a^3, ..., h_a^n\} \tag{2}$$

$$H^b = \{h_b^1, h_b^2, h_b^3, ..., h_b^m\} \tag{3}$$

3.3. Sequential Information Attention Mechanism

We present an innovative attention mechanism known as the sequential information attention mechanism (SIAM), specifically designed to exploit information interaction between sequence texts. Initially, we calculate the attention weight of the current text T_b to the historical text T_a:

$$w = softmax(H^a \cdot (H^b)^T) \tag{4}$$

Historical text H^a_{new} is updated with attention weights to capture differences between the current text and the historical text:

$$H^a_{new} = w \cdot H^a \tag{5}$$

For the updated H^a_{new}, we use the average pooling operation to obtain an enhanced global representation of historical text:

$$h^g = pooling(H^a_{new}) \tag{6}$$

3.4. Language Development Assessment

We concatenate the enhanced global representation h^g of historical text with the original sentence representation $h^{[CLS]}$ to obtain a text representation for classification:

$$h^c = concat(h^{[CLS]}, h^g) \tag{7}$$

where $h^{[CLS]}$ represents the semantic representation associated with the token "[CLS]" within the given sentence. The representation h^c is utilized as the sentence's overall feature representation, which is subsequently fed into a linear classifier with a softmax function. The predicted probabilities language development assessment of are:

$$y = softmax(W^T \cdot h^c + b) \tag{8}$$

where W and b are learnable parameters. The cross-entropy loss is employed to calculate the loss that penalizes the predicted class probability based on how far it is from the actual expected value. The cross-entropy loss function of language development assessment L_{ce} is defined as:

$$L_{ce} = -\sum_{j=1}^{2} e_j \cdot log(y_j) \tag{9}$$

where e is the one-hot encoding of the text's actual expected value. When $e = 1, 0$ means that the writing time of the text T_b is before the text T_a; otherwise, when $e = 0, 1$ means that the writing time of the text T_b is after the text T_a.

4. Experiments

4.1. Experimental Setup

All experimental procedures are conducted utilizing the NVIDIA A30 24-GB GPU. We utilize pytorch [14] and transformers [15] to build our models. Considering the similarity between the two languages, we only use the Italian BERT model (dbmdz/bert-base-italian-uncased), as we think it also contains a small amount of Spanish information. The feed-forward layer is initialized using weights drawn from a truncated normal distribution with a standard deviation of 2e-2, while the bias is initialized to zero. A fixed initial learning rate of 5e-5 is consistently applied across all experiments. The maximum sequence length is set to 512, representing the prescribed constraint on the number of tokens within a sentence. To optimize training, a warmup proportion of 1e-3 is implemented. The training episodes span 10 epochs with a batch size of 4.

4.2. Datasets

The datasets provided by EVALITA 2023 "Language Learning Development" task come from two samples, CItA [16] and COWS-L2H [9], where the number of training sets is 2394 and 1009 respectively. We perform data augmentation based on the datasets. Specifically, if essay 1 in sample 1 appears before essay 2, we describe it as (A_1,A_2,1), where '1' denotes the positive sample. While essay 2 in sample 2 appears before essay 3, we describe it as (A_2,A_3,1). Then we construct sample 3 based on the above two samples. In terms of sample 3, essay 1 appears before essay 3, which is defined as (A_1,A_3,1). In addition, we expand the negative samples based on the above positive samples, namely (A_2,A_1,0), (A_3,A_2,0), and (A_3,A_1,0), where '0' represents the the negative sample. The scales the augmented datasets for CItA and COWS-L2H are 5056 and 2042, respectively. In the training set, each positive sample can match a corresponding negative sample, so the number of positive and negative samples in the dataset is consistent. Ultimately, two datasets are combined to get a new training set.

In order to ensure the rationality of our strategy evaluation, we employ a 5-fold cross-validation methodology, which involves dividing the datasets into five distinct subsets to construct an ensemble model that exhibits enhanced generalization capabilities. More precisely, four of these subsets are assigned for training purposes, while the remaining subset is utilized for verification. The effective evaluation results of our strategies are derived by averaging the outcomes obtained from the five models.

4.3. Submission

We submit three results in total, namely the fine-tuned BERT model (Run 2), the fine-tuned BERT model with SIAM (Run 3), and the merge method (Run 1). The fine-tuned BERT model is to fine-tune directly on the BERT model based on the dataset. Concretely, the model in section 3 removes the sequential information attention mechanism. The merge method is the fusion of output probabilities of the fine-tuned BERT model (Run 2) and the fine-tuned BERT model with SIAM (Run 3).

4.4. Experimental results

Experimental results in the evaluation phase are shown in Table 1.

It can be seen that on the CItA test set, the BERT model achieves the best performance, F_{binary}, F_{macro} and the accuracy are 0.9338, 0.9315 and 0.9316 respectively, while the BERT model with SIAM has slightly declined. We deem that the impunity can be attributed to our methods being trained on two corpora simultaneously, to some extent, the information of the two corpora affects each other, sacrificing the performance of the CItA dataset in exchange for the improvement of the COWS dataset. Concerning the merge method, regardless of the CItA test set, the COWS test set or the combined test set, the strategy of model fusion is powerless, which has not brought effective improvement.

Our proposed sequential information attention mechanism has demonstrated substantial improvements in both the COWS test set and the combined test set. Specifically,

Table 1
Main Result.

Dataset	Metrics	BERT (Run 2)	BERT with SIAM (Run 3)	Merge (Run 1)
CItA	F_{binary}	0.9338	0.9260	0.9270
	F_{macro}	0.9315	0.9251	0.9250
	Acc	0.9316	0.9251	0.9251
COWS	F_{binary}	0.6201	0.6628	0.6306
	F_{macro}	0.6091	0.6391	0.6150
	Acc	0.6094	0.6406	0.6156
All	F_{binary}	0.7740	0.7883	0.7747
	F_{macro}	0.7669	0.7796	0.7669
	Acc	0.7671	0.7799	0.7671

on the COWS test set, the BERT model with SIAM outperforms the BERT model by 0.0427, 0.0300, and 0.0313 in the three indicators of F_{binary}, F_{macro}, and accuracy, respectively. Likewise, on the combined test set, the BERT model with SIAM gains consistent improvement of 0.0143, 0.0126, and 0.0128 in three metrics based on the BERT model.

5. Conclusion

The "Language Learning Development" task revolves around accurately predicting the sequential order of two essays authored by a single student. In the study, we first attempt to tackle the task leveraging a high-performing pre-trained language model, demonstrating the strong potential of pre-trained language models to solve the language development assessment task. Moreover, we present a novel attention mechanism, known as sequential information attention, designed to effectively capture and leverage the interaction of information within sequential texts. In the final evaluation stage, experimental results reveal the effectiveness of our proposed method, substantiating that sequential information attention contributes to tracking the evolution of language competence.

In the future, we will further try to focus on neural networks to extract language features suitable for language development assessment tasks, so as to further improve the performance of the model, driving advancements in the field of language development assessment.

Acknowledgments

This work was supported by the Guangdong Philosophy and Social Science Foundation (No. GD20CWY10), the National Social Science Fund of China (No. 22BTQ045), and the Science and Technology Program of Guangzhou (No.202002030227).

References

[1] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[2] S. A. Crossley, Linguistic features in writing quality and development: An overview, Journal of Writing Research 11 (2020) 415–443. URL: https://jowr.org/index.php/jowr/article/view/582. doi:10.17239/jowr-2020.11.03.01.

[3] K. Sagae, A. Lavie, B. MacWhinney, Automatic measurement of syntactic development in child language, in: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 197–204. URL: https://aclanthology.org/P05-1025. doi:10.3115/1219840.1219865.

[4] X. Lu, Automatic measurement of syntactic complexity in child language acquisition, International Journal of Corpus Linguistics 14 (2009) 3–28. doi:10.1075/ijcl.14.1.02lu.

[5] A. Miaschi, D. Brunato, F. Dell'Orletta, A nlp-based stylometric approach for tracking the evolution of l1 written language competence, Journal of Writing Research 13 (2021) 71–105. URL: https://www.jowr.org/index.php/jowr/article/view/778. doi:10.17239/jowr-2021.13.01.03.

[6] J. Hancke, D. Meurers, Exploring cefr classification for german based on rich linguistic modeling, 2013, pp. 54–56.

[7] S. Vajjala, K. Lõo, Automatic CEFR level prediction for Estonian learner text, in: Proceedings of the third workshop on NLP for computer-assisted language learning, LiU Electronic Press, Uppsala, Swe-

den, 2014, pp. 113–127. URL: https://aclanthology.org/W14-3509.

[8] I. Pilán, E. Volodina, Investigating the importance of linguistic complexity features across different datasets related to language learning, in: Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing, Association for Computational Linguistics, Santa Fe, New-Mexico, 2018, pp. 49–58. URL: https://aclanthology.org/W18-4606.

[9] A. Miaschi, S. Davidson, D. Brunato, F. Dell'Orletta, K. Sagae, C. H. Sanchez-Gutierrez, G. Venturi, Tracking the evolution of written language competence in L2 Spanish learners, in: Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, Seattle, WA, USA → Online, 2020, pp. 92–101. URL: https://aclanthology.org/2020.bea-1.9. doi:10.18653/v1/2020.bea-1.9.

[10] B. Bulté, A. Housen, Conceptualizing and measuring short-term changes in l2 writing complexity, Journal of Second Language Writing 26 (2014) 42–65. URL: https://www.sciencedirect.com/science/article/pii/S1060374314000666. doi:https://doi.org/10.1016/j.jslw.2014.09.005, comparing perspectives on L2 writing: Multiple analyses of a common corpus.

[11] K. Sagae, Tracking child language development with neural network language models, Frontiers in Psychology 12 (2021) 674402. doi:10.3389/fpsyg.2021.674402.

[12] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 6000–6010.

[14] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, Advances in neural information processing systems 32 (2019).

[15] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Huggingface's transformers: State-of-the-art natural language processing, arXiv preprint arXiv:1910.03771 (2019).

[16] A. Barbagli, P. Lucisano, F. Dell'Orletta, S. Montemagni, G. Venturi, CItA: an L1 Italian learners corpus to study the development of writing competence, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), European Language Resources Association (ELRA), Portorož, Slovenia, 2016, pp. 88–95. URL: https://aclanthology.org/L16-1014.

IUSS-NeTS at LangLearn: The role of morphosyntactic features in language development assessment

Matilde Barbini[1], Emma Zanoli[1] and Cristiano Chesi[1]

[1]University School for Advanced Studies IUSS Pavia - NeTS Lab

Abstract

This report presents the system developed by the IUSS-NeTS team for the LangLearn evaluation task of the 2023 edition of EVALITA [1], which focuses on tracking the linguistic development of native Italian speakers (L1) and second language learners of Spanish (L2). Our approach focuses on the integration of specific linguistic features and surprisal-based metrics for both Italian and Spanish datasets (CItA and COWS-L2H, respectively). With our tests the system achieves a F-score of 0.67 for the Italian test set and 0.75 for the Spanish test set. These results demonstrate the effectiveness of our methodology in capturing and analyzing the linguistic development of L1 Italian and L2 Spanish learners.

Keywords

language development assessment, linguistic features, surprisal, NLP

1. Introduction

LangLearn [2] is the first shared task on automatic language development assessment. It is defined as a binary classification problem where participants are asked to predict the relative order of two essays written by the same student, for both Italian L1 and Spanish L2. According to the resources employed to train the models, the task was separated into two sub-tasks:

- prediction obtained using only the training data made available for the task;
- prediction acquired utilizing also extra resources (e.g. additional data for the training phase).

Building upon prior research in the field [3, 4, 5, 6], our study aims to investigate the linguistic evolution of native Italian speakers and second language learners of Spanish. To achieve this, we employ a comprehensive approach that combines the extraction of linguistic features from the examined texts, encompassing syntactic, morphosyntactic, and lexical dimensions, with the addition of surprisal-based metrics. This multifaceted methodology allows us to gain deeper insights (especially from a transparent formal linguistic perspective) into the language proficiency development of L1 Italian and L2 Spanish learners. This work is focused on the first sub-task: our approach does not use any additional task-specific resources, although our use of language models in the calculation of surprisal-based metrics must be noted.

2. Data

In this section we describe the training and the test datasets for both Italian L1 and Spanish L2.

2.1. Training set

The CItA corpus (Corpus italiano di apprendenti L1), an archive of written works composed by the same students between the first and second year of lower secondary school, was the first corpus made available by the task's organizers. There are 1352 essays in the original corpus, written by 156 students; 2394 pairs of essays were assigned to task participants for the initial training session (note that some texts were used twice to create the pairs). The structure of the file containing the pairs was as follow: $essay_1$, $essay_2$, $order_1$, $order_2$. The essays' chronological placement varied between the first and second year of lower secondary school, contingent upon when they were written; meanwhile, the essay number denoted its sequential position within the given year.

The second dataset provided in the task was the COWS-L2H corpus (Corpus of written Spanish of L2 and Heritage speakers); it comprises 3498 brief

Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023)

matilde.barbini@iusspavia.it (M. Barbini); emma.zanoli@iusspavia.it (E. Zanoli); cristiano.chesi@iusspavia.it (C. Chesi)

0000-0002-0877-7063 (M. Barbini); 0000-0001-7116-9338 (E. Zanoli); 0000-0002-9421-8566 (C. Chesi)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper20.pdf

essays produced by second language learners attending a Spanish course at an American university, written in response to four different prompts; 1009 pairs were assigned to task participants for the initial training set. The structure of the file containing the pairs was the same as the Italian training set, however the columns labeled $order_1$ and $order_2$ contained chronological data to be interpreted as academic terms (quarters) and corresponding years. Each academic term spanned a specific time range: W denoted the time period from January to March, S from April to June, SU from July to September, and F from October to December[1].

2.2. Test set

The test set for the CItA corpus consisted of 278 texts organized into 307 pairs, whereas the test set for the COWS-L2H corpus comprised 471 texts organized into 320 pairs. It is important to note that, unlike in the train datasets, the essay identifying acronyms were presented in a random order.

3. System description

In this section we present our approach for the automatic tracking of Italian L1 and Spanish L2 language development. Our method can be summarized in a binary classification based on the combination of linguistic complexity features and surprisal-based metrics.

3.1. Linguistic complexity features

The aim of our approach was to evaluate the relevance of specific morphosyntactic features in the assessment of the linguistic competence improvement both in L1 and L2. With this in mind, we selected a set of robust Natural Language Processing (NLP) tools which allowed us to extract linguistic features relevant to the classification task and which was compatible with both Italian and Spanish. After an extensive evaluation of various tools (Tint[2], Profiling UD[3], READ-IT[4] and CTAP[5]), we decided to rely on CTAP: Common Text Analysis Platform [7]. The CTAP system supports the analysis of language complexity in a easy to use, platform indipendent, flexible and extendable environment [7]. The system was firstly available for German [8] and English [9], then it was adapted also to other languages [10]. In particular as it is stated in the paper which presented the Italian adaptation of the tool [10] CTAP is the most comprehensive linguistic complexity measurement tool for Italian and the only one allowing the comparison of Italian texts to multiple other languages within one tool.

3.2. Surprisal

We chose to increase the chances of success of our classification also by using surprisal-based metrics extracted from the constituent texts of the two LangLearn corpora: it has in fact been observed that the surprisal value of a word is associated with behavioral measures of processing complexity [11]. Surprisal-based metrics have already been used in various contexts, such as the analysis of pathological speech [12] and Italian linguistic studies [13]. When referring to surprisal, we are talking about one of the many possible log-transformed probability measures expressing the likelihood that a token follows a sequence of tokens (or it is placed in a certain position into a given sequence of tokens). We performed an extensive evaluation of the best models to be used in this case, testing 13 models on different linguistic structures (such as relatives, passives, questions, various typologies of errors) for both Spanish and Italian. One important trait we considered was the low tolerance for errors: whenever a morphosyntactic error is present, certain tools tend to normalize the ill-formed token eventually neglecting the relevant features that, in fact, indicate true mistakes. Expanding on previous studies [14], the incorporation of surprisal-based metrics was a deliberate effort to preserve the meaningfulness inherent in error occurrences. We set out to evaluate which model demonstrated the least inclination to normalize errors, examining the surprisal values generated by different models. We gave preference to those models that yielded higher values of surprisal when faced with various types of errors, indicating a recognition of the mistakes, over models that produced lower values leading to normalization of errors. After the evaluation we decided to use two BERT models for both Italian[6] and Spanish[7], available on HugginFace. In particular we calculated the surprisal for every word in every text giving to the two BERT models the entirety of the text as context. Upon detecting errors in the tokenization

[1] For a detailed description of the two corpora see [6] and [5].
[2] https://dh.fbk.eu/research/tint/
[3] http://www.italianlp.it/demo/profiling-ud/
[4] http://www.italianlp.it/demo/read-it/
[5] http://sifnos.sfs.uni-tuebingen.de/ctap/
[6] https://huggingface.co/dbmdz/bert-base-italian-cased
[7] https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased

Surprisal-based metrics	
text surprisal KURT	kurtosis
text surprisal MAX	maximum value
text surprisal MEAN	mean
text surprisal MEDIAN	median
text surprisal MIN	minimum value
text surprisal OUT F	outliers (frequency)
text surprisal OUT N	outliers (number)
text surprisal Q1	first quartile
text surprisal Q3	third quartile
text surprisal SKEW	skewness
text surprisal STD	standard deviation
text surprisal IQR	interquartile range

Table 1
12 surprisal-based metrics; every metric represents a statistic computed for the token surprisal distribution of every token in every text

performed by both models, we made the decision to introduce an additional preprocessing step which involved utilizing Stanford's parser Stanza[8] for tokenization of both the Italian and Spanish texts, prior to further analysis. Stanza is a Python natural language processing toolkit that supports 66 different human languages. Stanza's text analysis pipeline is language agnostic (in other words, language indipendent): the parser has been trained on 112 datasets, and it has been found that the same neural architecture generalizes efficiently and performs well across all languages [15]. Following the discussed methodology, we were able to calculate on both CItA and COWS-L2H corpora 12 statistical measures (table 1) based on the extraction of word surprisal, which were then added to the linguistic features previously extracted through the CTAP tool.

3.3. Classification

For the classification phase we decided to use Weka [16], since its rich selection of algorithms, ease of use, feature selection, evaluation capabilities and integration options make it a favorable choice for classification tasks[9]. Firstly we constructed ARFF files associating each text of the two corpora with the attributes consisting of the features extracted through CTAP and the 12 surprisal-based metrics. We set our classification in this way: give 1 if it is true that the $essay_1$ (i.e. the first essay of the pair) was written before $essay_2$ (i.e. the second essay of the pair), otherwise give 0. We pre-processed our files and we performed attribute selection on the extracted linguistic features and surprisal-based metrics: we calculated the correlation between each attribute and the output variable and selected only those attributes that had a moderate-to-high positive or negative correlation (close to -1 or 1), discarding the ones with a low correlation (value close to 0). This made it possible to eliminate attributes that were not significant in our analysis. For the Italian language we had initially 266 attributes which became 40 after the attribute selection phase, while for the Spanish language we had 400 attributes which became 75. To make a good use of the two training datasets, we performed the analysis with the cross-validation method (also known as k-fold validation), which is particularly useful to reduce the overfitting problem. To obtain the most comprehensive view we then evaluated several possible classification algorithms offered by Weka. We finally observed that the best performing algorithms were the Logistic regression, J48 and Random Forest, with the last one being the absolute best and thus our choice for the final analysis. Random Forest is a class of ensamble methods which aggregates the prediction of different classifiers to improve accuracy; Weka makes available the implementation of the algorithm described in [18].

4. Results

The final results yielded an F-score of 0.67 for the Italian test set and 0.75 for the Spanish test set. Weka facilitated the extraction of diverse informative measures regarding the classification performance (detailed performance metrics during the 10-fold cross-validation phase can be found in tables 2 and 3).

It is crucial to interpret all this information collectively to gain an accurate feedback on the classification. Among these measures, the ROC Area stands out as a highly informative index. It gauges

[8] https://stanfordnlp.github.io/stanza/
[9] For an extensive discussion of Weka's features see [17].

Detailed accuracy by class								
TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,749	0,304	0,716	0,749	0,732	0,446	0,791	0,781	0.0
0,696	0,251	0,731	0,696	0,713	0,446	0,791	0,786	1.0
Weighted average								
0,723	0,278	0,723	0,723	0,723	0,446	0,791	0,784	

Table 2
Detailed metrics of the cross-validation phase on the Italian training dataset.

Detailed accuracy by class								
TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,734	0,263	0,735	0,734	0,734	0,471	0,818	0,825	0.0
0,737	0,266	0,736	0,737	0,736	0,471	0,818	0,824	1.0
Weighted average								
0,735	0,265	0,735	0,735	0,735	0,471	0,818	0,824	

Table 3
Detailed metrics of the cross-validation phase on the Spanish training dataset.

the model's overall performance by calculating the area under the curve in a graph plotting the TP (True Positive) rate (y-axis) against the FP (False Positive) rate (x-axis), as the classification decision threshold changes. An optimal classifier exhibits ROC area values close to 1, while 0.5 is akin to a random guess (similar to a Kappa statistic of 0).

5. Analysis and discussion of the results

The intuition behind our approach was that, in tracing the linguistic evolution, we should observe a change/maturation in the syntactic, morphosyntactic and lexical characteristics of the texts examined. Nevertheless, it was necessary to make further considerations and formulate precise hypotheses: if one could expect an increase in lexical sophistication and variety for both the Italian and Spanish corpus texts, it was however doubtful what phenomena we would have witnessed from the syntactic point of view. Even though we have evidence supporting the theory that, in normal acquisition, functionally higher features (such as those involved in complementation, question or relative clauses formation) are developed at later stages of acquisition [19], we wanted to assess as many linguistic features as possible in order to map a large number of phenomena and characteristics of the students' essays. Let us now examine in more detail which attributes were selected for classification.

5.1. Features selected for the Italian dataset

Out of the 40 attributes chosen for the Italian dataset, 11 were specifically associated with the density of different part-of-speech categories. These included the density of conjunctions, subjunctive mode verbs, pronouns of various types, and auxiliary verbs. Additionally, 2 quantitative features were considered, namely the count of sentences and the average sentence length measured in tokens. 15 attributes pertained to the quantity of various types of syntactic constituents identified within the texts. Furthermore, 12 attributes were focused on assessing the sophistication, refinement, and variety of the lexicon employed. Lastly, only 1 surprisal-based metric was selected, which involved the identification of outliers (essentially the count of anomalous values that significantly deviated from the remaining observations)[10].

5.2. Features selected for the Spanish dataset

Out of the 75 attributes chosen for the Spanish dataset, 6 were specifically associated with the density of various part-of-speech categories. These included the density of verbs, adverbs, adjectives, and others. Additionally, 9 attributes were quantitative features, encompassing measurements such as the mean sentence length in letters, the number of tokens, the standard deviation of sentence length in syllables and letters, and the count of tokens with more than two syllables, among others. A total of 29 features were utilized to track morphological and syntactic complexity. These features encompassed the count of various syntactic constituents, as well as specific indicators such as the number of relative clauses, prepositional phrases per clause, and the

[10]For a list of the selected features for Italian see https://drive.google.com/file/d/1FCIn1DsawUR2YYg3L685q09SZ4NaEtnX/view?usp=sharing

occurrence of passive structures. Moreover, 24 features were dedicated to assessing the sophistication, refinement, and variety of the lexicon employed. 2 additional indices focused on textual cohesion, particularly examining lemma overlaps. Lastly 5 metrics based on surprisal were selected. These metrics included the interquartile range (i.e., the spread of the middle half of the data), the median, the minimum value, the third quartile (i.e., the middle value between the dataset's median and the highest value), and the standard deviation [11].

5.3. Concluding remarks

As a general observation, it is evident that the performance achieved on the Spanish dataset surpasses that of the Italian dataset. This discrepancy can potentially be attributed to multiple factors. One plausible explanation is the larger number of available attributes for the Spanish language compared to Italian. Additionally, it is noteworthy that the acquisition of a second language (L2) exhibits more significant variation and development within a shorter temporal period compared to a first language (L1). The elevated significance of surprisal-based metrics in relation to the texts from the Spanish L2 corpus can also be rationalized by a higher occurrence of errors and inconsistency/incoherence within the texts during the initial stages of L2 language learning. It is our belief that the outcomes of this study should be interpreted as motivation to further explore the linguistic characteristics of texts, with a particular emphasis on the development of additional indices pertaining to linguistic complexity and information theory. This pursuit aims to gain a more comprehensive understanding of the similarities and differences in the progression of L1 and L2 languages.

Acknowledgments

We are thankful to the colleagues of NeTS lab at IUSS Pavia who discussed with us the preliminary results of these experiments. This research has been partially founded by PON Governance 2014-2020: Next Generation UPP Project - CUP D19J22000240006.

[11] For a list of the selected features for Spanish see https://drive.google.com/file/d/16ugOhuL6P8eW1v-C7YaKuVhL8RmifRRv/view?usp=sharing

References

[1] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[2] C. Alzetta, D. Brunato, F. Delll'Orletta, A. Miaschi, K. Sagae, C. H. Sánchez-Gutiérrez, G. Venturi, Langlearn at evalita 2023: Overview of the language learning development task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[3] S. Vajjala, Automated assessment of non-native learner essays: Investigating the role of linguistic features, International Journal of Artificial Intelligence in Education 28 (2016) 79–105.

[4] Z. Weiss, D. Meurers, Analyzing linguistic complexity and accuracy in academic language development of German across elementary and secondary school, in: Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, Florence, Italy, 2019, pp. 380–393. URL: https://aclanthology.org/W19-4440. doi:10.18653/v1/W19-4440.

[5] A. Miaschi, S. Davidson, D. Brunato, F. Dell'Orletta, K. Sagae, C. H. Sanchez-Gutierrez, G. Venturi, Tracking the evolution of written language competence in L2 Spanish learners, in: Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, Seattle, WA, USA → Online, 2020, pp. 92–101. URL: https://aclanthology.org/2020.bea-1.9. doi:10.18653/v1/2020.bea-1.9.

[6] A. Miaschi, D. Brunato, F. Dell'Orletta, A nlp-based stylometric approach for tracking the evolution of l1 written language competence, Journal of Writing Research 13 (2021) 71–105. URL: https://www.jowr.org/index.php/jowr/article/view/778. doi:10.17239/jowr-2021.13.01.03.

[7] X. Chen, D. Meurers, CTAP: A web-based

tool supporting automatic complexity analysis, in: Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC), The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 113–119. URL: https://aclanthology.org/W16-4113.

[8] J. Hancke, S. Vajjala, D. Meurers, Readability classification for German using lexical, syntactic, and morphological features, in: Proceedings of COLING 2012, The COLING 2012 Organizing Committee, Mumbai, India, 2012, pp. 1063–1080. URL: https://aclanthology.org/C12-1065.

[9] S. Vajjala, D. Meurers, Assessing the relative reading level of sentence pairs for text simplification, in: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Gothenburg, Sweden, 2014, pp. 288–297. URL: https://aclanthology.org/E14-1031. doi:10.3115/v1/E14-1031.

[10] N. Okinina, J.-C. Frey, Z. Weiss, CTAP for Italian: Integrating components for the analysis of Italian into a multilingual linguistic complexity analysis tool, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 7123–7131. URL: https://aclanthology.org/2020.lrec-1.880.

[11] N. J. Smith, R. Levy, The effect of word predictability on reading time is logarithmic, Cognition 128 (2013) 302–319. URL: https://www.sciencedirect.com/science/article/pii/S0010027713000413. doi:https://doi.org/10.1016/j.cognition.2013.02.013.

[12] N. Rezaii, J. Michaelov, S. Josephy-Hernandez, B. Ren, D. Hochberg, M. Quimby, B. Dickerson, A computational approach for measuring sentence information via surprisal: theoretical implications in nonfluent primary progressive aphasia, medRxiv (2022). URL: https://www.medrxiv.org/content/early/2022/11/29/2022.11.25.22282630. doi:10.1101/2022.11.25.22282630. arXiv:https://www.medrxiv.org/content/early/2022/11/29/2022.11.25.22282630.full.pdf.

[13] J. A. Michaelov, B. K. Bergen, Do language models make human-like predictions about the coreferents of italian anaphoric zero pronouns?, 2022. arXiv:2208.14554.

[14] G. Kharkwal, S. Muresan, Surprisal as a predictor of essay quality, in: Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, Baltimore, Maryland, 2014, pp. 54–60. URL: https://aclanthology.org/W14-1807. doi:10.3115/v1/W14-1807.

[15] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, C. D. Manning, Stanza: A python natural language processing toolkit for many human languages, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 101–108. URL: https://aclanthology.org/2020.acl-demos.14. doi:10.18653/v1/2020.acl-demos.14.

[16] E. Frank, M. A. Hall, G. Holmes, R. Kirkby, B. Pfahringer, I. H. Witten, Weka: A machine learning workbench for data mining., Springer, Berlin, 2005, pp. 1305–1314. URL: http://researchcommons.waikato.ac.nz/handle/10289/1497.

[17] D. Merlini, M. Rossini, Text categorization with weka: A survey, Machine Learning with Applications 4 (2021) 100033. URL: https://www.sciencedirect.com/science/article/pii/S2666827021000141. doi:https://doi.org/10.1016/j.mlwa.2021.100033.

[18] L. Breiman, Random forests, Machine learning 45 (2001) 5–32.

[19] N. Friedmann, A. Belletti, L. Rizzi, Growing trees: The acquisition of the left periphery, Glossa: a journal of general linguistics 6 (2021). URL: https://www.glossa-journal.org/article/id/5877/. doi:10.16995/glossa.5877.

bot.zen at LangLearn: regressing towards interpretability

Egon W. Stemle[1,2,*], Martina Tebaldini[3,1], Francesca Bonanni[3,1], Filippo Pellegrino[3,1], Paolo Brasolin[1], Greta H. Franzini[1], Jennifer-Carmen Frey[1], Olga Lopopolo[1,4] and Stefania Spina[1,4]

[1]Institute for Applied Linguistics, Eurac Research, Viale Druso, 1, 39100 Bolzano (BZ), Italy
[2]Faculty of Informatics, Masaryk University, Czech Republic
[3]University of Bolzano, Italy
[4]Università per Stranieri di Perugia, Italy

Abstract

This article describes the `bot.zen` system that participated in the Language Learning Development (LangLearn) shared task of the EVALITA 2023 campaign. We developed a simple machine learning system with good interpretability for later use, and used the shared task as an opportunity to provide Master's students with hands-on training and practical experience in NLP.

Keywords

system description, langlearn, evalita, shared task, regression, MALT-IT2, bot.zen

1. Introduction

There has been an increasing interest in using Natural Language Processing (NLP) tools and machine learning techniques to analyse writing development in first (L1) and second language (L2) acquisition settings. The topic has been explored in Second Language Acquisition (SLA), Learner Corpus Research (LCR) (e.g. [1]), Corpus Linguistics and in writing development research (e.g. [2]), and its goal is to understand how specific features can reflect writing quality and development.

The analysis of language learner data typically spans sociolinguistic metadata (information about the author), linguistic data (information extracted from the text) and textual metadata (information about the text). According to [3], metadata such as reading time, geographic factors, and parents' occupation level can have an impact on language skill development, whereas [4] finds writing quality and development to be influenced by both text length and linguistic features including lexical density, diversity and sophistication, as well as syntactic complexity and text cohesion. Finally, a text usually includes metadata such as the author, the date of creation, the context in which it was written, and a language proficiency rating. This contextual information enhances the overall understanding of the content. All of these research strands can support NLP applications for writing evaluation and assessment, including automatic essay scoring, automatic writing evaluation systems, and automatic classification of text difficulty for learners. (For an in-depth overview and additional references, see [4].)

At EVALITA 2023 [5], the Language Learning Development (LangLearn) shared task (ST) on automatic language development assessment [6] consisted in predicting the relative order of two essays written by the same student. More specifically, the texts provided were in Italian and Spanish, and came with only a very limited set of metadata. We participated in this ST to acquire experience with this type of data, and as an opportunity to involve and train Master's students from the University of Bolzano[1] in NLP scientific work through practical internships.

Our system relies only on the data provided for the ST, generates explicit information about students' progress out of implicit information in the data and uses regression without Large Language Models (LLMs) or Neural Networks (NNs) with features from an external tool specifically designed for Italian texts. As a result, our system performed well on Italian but poorly on Spanish data.

The rest of the paper is organised as follows: Section 2 describes the system design and implementation; Section 3 describes our experiments and results; and Section 4 concludes with a short discussion.

[1]https://www.unibz.it/en/faculties/education/master-applied-linguistics/

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
*Corresponding author.
egon.stemle@eurac.edu (E. W. Stemle); martina.tebaldini@student.unibz.it (M. Tebaldini); francesca.bonanni@student.unibz.it (F. Bonanni); filippo.pellegrino@student.unibz.it (F. Pellegrino); paolo.brasolin@eurac.edu (P. Brasolin); greta.franzini@eurac.edu (G. H. Franzini); jennifercarmen.frey@eurac.edu (J. Frey); olga.lopopolo@eurac.edu (O. Lopopolo); stefania.spina@eurac.edu (S. Spina)
https://iiegn.eu (E. W. Stemle)
0000-0002-7655-5526 (E. W. Stemle); 0000-0003-2471-7797 (P. Brasolin); 0000-0003-1159-5575 (G. H. Franzini); 0000-0002-7008-6394 (J. Frey); 0000-0003-0997-367X (O. Lopopolo); 0000-0002-9957-3903 (S. Spina)
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper21.pdf

2. System Design and Implementation

Our objective was to develop a simple machine learning system with good interpretability. Therefore, we prioritised a simple design that could provide transparent explanations for its decision-making process over complex implementation and high predictive performance.

2.1. Data Pre-processing

In a first processing step, we restructure the given ST data, which provides essay ids with their respective time of writing in tabular format, as shown in Figure 1.

essay_1	essay_2	order_1	order_2
5074	4666	1_1	1_2
5074	4948	1_1	1_4
5074	4872	1_1	1_5
5074	4937	1_1	2_1
5074	5128	1_1	2_2
5074	4212	1_1	2_3
5074	5361	1_1	2_4

Figure 1: Excerpt for one author from the 4-column `Training_CItA.tsv` file with the first two columns (`essay_1,2`) representing essay ids and the last two columns (`order_1,2`) their respective time of writing (`YEAR_Essay-Number`).

We reconstruct the data for individual authors by collecting all pairs with overlapping essays and reordering them based on the provided information, as shown in Figure 2. This new data structure contains an *absolute position* for the time of writing, which eventually becomes the target variable to learn.

With this remodelled dataset, we transform a classification task into a regression task: rather than deciding whether one text was written earlier than the other in a binary fashion, we predict the absolute position of the texts which, in turn, can be used to calculate the order in which the two essays were written.

2.2. Feature Extraction

We use spaCy [7] and MALT-IT2 [8] in order to transform the raw input data into a meaningful set of informative features, as they provide easy-to-use and reliable feature extraction methods.

2.2.1. spaCy

spaCy is an open-source NLP library in Python providing tools for many tasks and pre-trained models for several languages, including Italian and Spanish[2].

order:essay	abs.position
1_1:5074	1
1_2:4666	2
1_4:4948	4
1_5:4872	5
2_1:4937	7
2_2:5128	8
2_3:4212	9
2_4:5361	10

Figure 2: Excerpt from our restructured `Training_CItA.tsv` file representing all essays written by one author with their time of writing and inferred abs(olute) positions. `2_4:5361` means that `essay:5361` was the fourth essay to be written in `YEAR:2` by this author, which corresponds to the overall tenth position for this author relative to all other authors (because the author missed `order` positions `1_3` and `1_6` in the first year).

After tokenisation, we collect 1- to 3-grams of the word forms and the part-of-speech tags. We additionally collect 2-grams of the morphological analyses of the words and 1-grams of a word's dependency relation. Overall, this amounts to roughly 17,000 features per document.

2.2.2. MALT-IT2

MALT-IT2 is an automatic classification system for measuring the complexity level of Italian texts, and it is based on experiments that compared ten machine learning models on a dataset of 692 texts and 139 linguistic features. The features were divided into the following six broad categories [9]:

1. Raw Text Features are the most elementary features and they are based on simple counting procedures executed on the tokenised text.
2. Lexical Features are computed by considering: (i) lemmas, POS tags and morphological annotations; (ii) external resources for the Italian language, for example, the vocabulary rate in the "new basic Italian vocabulary".
3. Morphological Features consider the morphological complexity index (MCI) computed for two word classes: verbs and nouns.
4. Morpho-Syntactic Features are computed on the basis of POS tagging, morphological analysis, and syntactic parsing.
5. Syntactic Features reflect the main characteristics and the structure of the syntactic constituents and the dependency relations of the sentences in a text.

[2]We use the `it_core_news_lg` and `es_core_news_lg` models.

6. Discursive Features take into account the cohesive structure of a text.

MALT-IT2 has to be invoked externally to process text files into a comma-separated values (CSV) file containing one line per document within its feature space; the CSV file is subsequently ingested by our system without any additional interaction or knowledge of MALT-IT2. This means that we can swap out MALT-IT2 with a different system or add another system capable of producing a document-feature-matrix in CSV format.

2.2.3. CTAP

We also experimented with a version of the Common Text Analysis Platform (CTAP) [10] adapted for Italian text [11]. Much like MALT-IT2, CTAP is a linguistic complexity measurement tool offering various statistics and features to analyse text complexity in terms of length, lexical, syntactic and morpho-syntactic aspects. Unfortunately, we encountered some problems while processing the entire dataset. Very short texts, for instance, caused CTAP to end prematurely with no error message, leaving us with no choice but to exclude CTAP features from our system. CTAP is capable of producing a document-feature-matrix in CSV format and could have been easily integrated into our system.

2.3. Processing Pipeline

Our data processing pipeline has been implemented in Python and makes use of the pandas and scikit-learn libraries[3].

pandas [12] is an open-source library for data manipulation and analysis that integrates well with other libraries in the Python ecosystem, making it a versatile tool for data analysis and preparation.

Our system uses pandas for internal data representations, manipulations and calculations during data preprocessing (Section 2.1) and the processing of CSV files.

scikit-learn [13] is an open-source machine learning library for Python providing a wide range of algorithms and tools for various tasks, including classification and dimensionality reduction. With a user-friendly and consistent interface, extensive documentation and an established user base, scikit-learn makes it easy to implement machine learning workflows.

Our system uses scikit-learn for the main processing, as illustrated in Figure 3.

The processing pipeline requires a document feature matrix that represents all texts as vectors in our feature space. This space is the combination (concatenation) of all different tools after feature extraction (Section 2.2), totalling around 17,200 features. To standardise the data, we use the `StandardScaler()`, which removes the mean and scales it to unit variance. We also reduce the linear dimensions using the `TruncatedSVD()` method[4]. As a results, our processed dataset consists of 125 features. Finally, we use the `HistGradientBoostingRegressor()` for learning, which is an ensemble method[5].

```
pipeline = Pipeline(steps=[
    ('combined_features',
        combined_features),
    ('scaler', StandardScaler()),
    ('redux', TruncatedSVD(125)),
    ('estimator', HGBoostingRegressor(
        loss='squared_error'))
])
```

Figure 3: The core of the processing pipeline. From top to bottom, the combined features are scaled, reduced and processed with only minimal parameterisation.

In order to use our system for the ST, we perform data post-processing. We convert the output of our regression model for individual texts into a binary label for pairs of texts that indicated which of the two was written first.

2.4. Optimisation

The different parts of our system were optimised towards our target variable (absolute position) via an ad-hoc grid search in 3-fold cross validation (CV) runs.

The parts we optimised were: the types of spaCy information to collect[6]; n-gram ranges and the minimum document frequencies for the spaCy collectors; the type of dimensionality reduction[7] and the number of dimensions to use; the regression algorithm to use[8].

3. Experiments and Results

3.1. Shared Task (ST)

The Language Learning Development (LangLearn) ST [6] consisted in predicting the relative order of two essays: given a randomly ordered pair (Essay 1, Essay 2) written

[3] We used `Python 3.8.16`, `pandas 2.0.1`, `scikit-learn 1.2.2`, and `spacy 3.5.3 + it-core-news-lg 3.5.0` for processing.

[4] We perform feature reduction to remove noise or irrelevant information, and highlight important aspects of the data, enabling the model to make more accurate predictions.

[5] Ensemble methods combine and aggregate predictions of multiple models to improve predictive performance.

[6] `token.text`, `token.lemma_`, `token.pos_`, `token.morph`, `token.dep_`

[7] `PCA()`, `TruncatedSVD()`

[8] `DecisionTreeRegressor()`, `SVR()`, `KernelRidge()`, `HistGradientBoostingRegressor()`

by the same student, the task was to predict whether Essay 1 had been written before Essay 2.

3.2. Shared Task Data

The LangLearn ST data contains essays from two different corpora, namely CItA [14] and COWS-L2H [15], with texts in Italian and Spanish, respectively.

Training data includes information on pairs of texts written by the same student at different times. Each entry represents the sequence of two essays, and by considering multiple entries with overlapping text-ids we are able to recreate the sequence of all texts for each student (see Section 2.1). The data also contains the texts themselves but no additional (meta)information beyond this.

CItA The CItA corpus (Corpus Italiano di Apprendenti L1) is a collection of Italian essays written by students learning their first language in seven different lower secondary schools in Rome over the course of two years (2012-2013 and 2013-2014). The students were asked to write different types of essays, namely reflexive, narrative, descriptive, expository and argumentative. The ST data contains 834 of the total 1,352 essays written but does not provide any information about the type of text.

COWS-L2H The COWS-L2H corpus (Corpus of Written Spanish of L2 and Heritage Speakers) is a collection of texts created by students of Spanish as a second language enrolled at a North American university. The students were asked to write multiple compositions at different times throughout the academic quarters, and the essays were collected over the course of two years, from 2017 to 2020. The essays were written by the same students, and the ST data contains 1,426 of the original 3,498 essays.

3.3. Results

The performance of our system on the two datasets (as reported by the ST organisers) was:

		acc	f-score
CItA	bot.zen	0.83	0.84
	best	0.93	0.93
	baseline	0.55	0.55
		acc	f-score
COWS-L2H	bot.zen	0.50	0.52
	best	0.75	0.75
	baseline	0.66	0.66

The baseline scores were calculated by training a LinearSVM using the number of tokens per document and the Type-Token-Ratio (TTR) of the first 100 tokens in each document as input features.

3.4. Analysis

We also analysed the CItA-part of the ST dataset independent of our system's performance. To this end, we used the original data with texts in *Set 1* always written before texts in *Set 2*. We then used CTAP to calculate feature values for all texts in both sets. Afterwards, we conducted a (paired) t-test to detect features that differed in their means (as a starting point for later research).

We found some evidence that *Set 1* had a higher number of 'basic vocabulary' words, whereas *Set 2* had a higher number of imageability words. *Set 1* also had higher TTR and HDD (Hypergeometric Distribution D) measures, but since *Set 2* generally had longer texts, length effects certainly come into play [16]. Also, *Set 1* used more auxiliary verbs, possibly due to a higher presence of past participle verbs. The use of connectives was higher in *Set 2*, especially for additive and consequence connectives. The number of dependent clauses per sentence did not differ significantly between the sets. Finally, *Set 2* contained more sentences and more punctuation marks but sentence length remained constant.

4. Discussion

Our system (see Section 2) was relatively simple. Neither LLMs nor recurrent neural networks (RNNs) were integrated, nor did we use any data other than those provided by the organisers. While our results for the Italian data were satisfactory, we performed very poorly on the Spanish data, as expected: MALT-IT2, our main processing component, was designed for Italian texts only, which had a negative impact on our system when processing Spanish data, and despite the baseline system information also being encoded in our features, the presence of too much irrelevant data hampered the overall performance.

Nevertheless, the ST served as a great opportunity for Master's students to gain practical project work experience: running into all-too-common data processing, encoding and decoding difficulties whilst navigating the intricacies of understanding, analysing and evaluating the data for the task at hand. With the help of the literature suggestions provided by the organisers, the students were able to develop relevant ideas and provide target-oriented answers to emerging questions. Although the internship was only 150 hours long and did not include the implementation of a functional application[9], the students had the opportunity to familiarise themselves with the crucial stages of a scientific project, documenting all steps into a project report, which was partially incorporated in this paper.

[9]Eurac Research took over the task of implementing a functional application.

Acknowledgments

We would like to thank our colleagues Arianna Bienati, Francesco Fernicola and Lionel Nicolas for their support during the project.

References

[1] S. A. Crossley, D. S. McNamara, Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners, Journal of Second Language Writing 26 (2014) 66–79. doi:10.1016/j.jslw.2014.09.006.

[2] P. Durrant, M. Brenchley, L. McCallum, Understanding Development and Proficiency in Writing: Quantitative Corpus Linguistic Approaches, 1st ed., Cambridge University Press, 2021. doi:10.1017/9781108770101.

[3] A. Barbagli, F. Dell'Orletta, G. Venturi, P. Lucisano, S. Montemagni, Il ruolo delle tecnologie del linguaggio nel monitoraggio dell'evoluzione delle abilità di scrittura: primi risultati (2015) 105–123. URL: https://journals.openedition.org/ijcol/326.

[4] S. A. Crossley, Linguistic features in writing quality and development: An overview, Journal of Writing Research 11 (2020) 415–443. doi:10.17239/jowr-2020.11.03.01.

[5] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, EVALITA 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for Italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[6] C. Alzetta, D. Brunato, F. Delll'Orletta, A. Miaschi, K. Sagae, C. H. Sánchez-Gutiérrez, G. Venturi, LangLearn at EVALITA 2023: Overview of the language learning development task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[7] M. Honnibal, I. Montani, spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, 2017. URL: https://spacy.io/.

[8] L. Forti, G. Grego Bolli, F. Santarelli, V. Santucci, S. Spina, MALT-IT2: A new resource to measure text difficulty in light of CEFR levels for Italian L2 learning, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 7204–7211. URL: https://aclanthology.org/2020.lrec-1.890.

[9] V. Santucci, F. Santarelli, L. Forti, S. Spina, Automatic classification of text complexity, Applied Sciences 10 (2020) 7285. doi:10.3390/app10207285.

[10] X. Chen, D. Meurers, CTAP: A web-based tool supporting automatic complexity analysis, in: Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC), The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 113–119. URL: https://aclanthology.org/W16-4113.

[11] N. Okinina, J.-C. Frey, Z. Weiss, CTAP for Italian: Integrating components for the analysis of Italian into a multilingual linguistic complexity analysis tool, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 7123–7131. URL: https://aclanthology.org/2020.lrec-1.880.

[12] The pandas development team, pandas-dev/pandas: Pandas 2.0.1, Zenodo, 2020. URL: https://pandas.pydata.org/. doi:10.5281/zenodo.3509134.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830. URL: https://scikit-learn.org/.

[14] A. Barbagli, P. Lucisano, F. Dell'Orletta, S. Montemagni, G. Venturi, CItA: an L1 Italian learners corpus to study the development of writing competence, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation, European Language Resources Association, Portorož, Slovenia, 2016, pp. 88–95. URL: https://aclanthology.org/L16-1014.

[15] A. Yamada, S. Davidson, P. Fernández-Mira, A. Carando, K. Sagae, C. Sánchez-Gutiérrez, COWS-L2H: A corpus of Spanish learner writing, Research in Corpus Linguistics 8 (2020) 17–32. doi:10.32714/ricl.08.01.02.

[16] M. Stills, Language Sample Length Effects on Various Lexical Diversity Measures: An Analysis of Spanish Language Samples from Children, Technical Report, Portland State University, 2016. doi:10.15760/honors.250.

A. Online Resources

- The bot.zen system for the EVALITA 2023 LangLearn shared task (on GitHub)

HaSpeeDe3 at EVALITA 2023: Overview of the Political and Religious Hate Speech Detection task

Mirko Lai[1,*], Fabio Celli[2], Alan Ramponi[3], Sara Tonelli[3], Cristina Bosco[1] and Viviana Patti[1]

[1]*Università degli Studi di Torino, Torino, Italy*
[2]*Maggioli s.p.a., University of Trento, Trento, Italy*
[3]*Fondazione Bruno Kessler (FBK), Trento, Italy*

Abstract

The Hate Speech Detection (HaSpeeDe3) task is the third edition of a shared task on the detection of hateful content in Italian tweets. It differs from the previous editions while maintaining continuity in analysing and contrasting hate speech (HS) on social media. While HaSpeeDe and HaSpeeDe2 were focused on HS against immigrants, Muslims and Roms, HaSpeeDe3 explores hate speech in strong polarised debates, concerning in particular politics and religion. It is articulated in two different tasks: A) In-domain political hate speech detection and B) Cross-domain hate speech detection about political and religious tweets. Task A consists in two different subtasks for which participants i) can only use the provided textual content of the tweet, or ii) can additionally employ contextual information about the tweet and its author. In Task B, that consists in two subtasks, participants are allowed to use any kind of external data for detecting hate speech in tweets about i) politics and ii) religion. Six teams from both academia and industry participated in the evaluation, with a total of 13 submitted runs for Task A and 16 for Task B.

Keywords

Hate speech detection, social media analysis, polarised debates, political hate speech, religious hate speech, shared task

1. Introduction and Motivation

Social media play an important role in public debates, especially concerning politics. On the one hand, political leaders use social media as a vehicle for political and electoral propaganda. On the other hand, they provide news to a significant part of the population that takes part in the discussion, supporting or criticising political decisions [1, 2]. Social media are also the place where debates on sensitive topics, such as religious beliefs and practices, are rather common and sometimes are intertwined with public discussions on political matters.

Unfortunately, such discussions often trigger verbal aggressions [3], especially after some polarising events in Europe and beyond such as Brexit [4], the Covid-19 pandemics [5] and the Russo-Ukrainian conflict [6]. Aggressions and online hate are exacerbated by the ideological segregation present on social media, where social homophily, as well as personalising and recommending algorithms, facilitate the creation of *echo chambers* and *filter bubbles* [7, 8]. The "others" are frequently targeted because of characteristics such as gender, sexual orientation, ethnicity, and religion [9, 10, 11].

In the last years, to address these problems posed by the widespread use of abusive language online, the NLP community has focused on the detection of hate speech [12] and the analysis of online debates [13, 14]. In particular, many researchers have worked on systems to detect offensive language against specific vulnerable groups, e.g., women, immigrants, LGBTQ+ community, among others [11, 15, 16, 17]. An under-researched – yet important – area of investigation is anti-politics hate, i.e., hate speech against politicians, policy makers and laws at any level (national, regional and local). While anti-policy hate speech has been addressed in Arabic [18] and German [19], most European languages have been under-researched. As regards religious hate, instead, annotated corpora have been created for English, Arabic, Bengali, French, Portuguese, and Italian, among others (for an overview of works, see [15] and [20]). However, none of them share contextual information about the authors of the tweets, neither about their social media network, although religious self-identification may lead to hard conflict with the members of other worships.

For this shared task organised within EVALITA 2023 [21], we introduce a new corpus, called PolicyCorpusXL, containing Italian tweets related to political topics, where

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
*Corresponding author.
mirko.lai@unito.it (M. Lai); Fabio.Celli@maggioli.it (F. Celli); alramponi@fbk.eu (A. Ramponi); satonelli@fbk.eu (S. Tonelli); cristina.bosco@unito.it (C. Bosco); viviana.patti@unito.it (V. Patti)
http://www.di.unito.it/~lai/ (M. Lai); https://dh.fbk.eu/author/alramponi/ (A. Ramponi); https://dh.fbk.eu/author/sara/ (S. Tonelli); http://www.di.unito.it/~bosco/ (C. Bosco); http://www.di.unito.it/~patti/ (V. Patti)
0000-0003-1042-0861 (M. Lai); 0000-0002-7309-5886 (F. Celli); 0000-0002-4305-2404 (A. Ramponi); 0000-0001-8010-6689 (S. Tonelli); 0000-0002-8857-4484 (C. Bosco); 0000-0001-5991-370X (V. Patti)
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper22.pdf

hateful messages have been manually annotated. This corpus is an extension of PolicyCorpus [22]. We selected Twitter as the source of data and Italian as the target language because Italy has, at least since the elections in 2018, a large audience that pays attention to hyper-partisan sources on Twitter. These users are prone to produce and retweet messages of hate against policy-making [23]. We also provide the Italian portion of the ReligiousHate dataset [20] as a test set, in which hateful tweets concerning Christianity, Islam and Judaism have been manually labeled. Our goal is to test the in-domain performance of systems for political hate speech detection, as well as the out-of-domain performance on a test set about religion.

2. Definition of the Task

HaSpeeDe3 focuses on detecting hate speech in strong polarised debates on social media, in particular debates on Twitter about political and religious topics. With this task, we invite participants to explore not only features based on the textual content of the tweet, but also features based on contextual information such as metadata that describe both the tweet and the author, or information about the social media community of the participants of the debate.

We propose two tasks, A and B, that in the rest of the paper will be referred also as in-domain and cross-domain tracks. Both tasks aim at tackling binary classification problems, and thus participants' systems have to predict whether a tweet contains hatred or not. Each task consists of two subtasks:

- **Task A – (In-domain) political hate speech detection**: a binary classification task aimed at determining whether a message contains hate speech or not. The task is based on the PolicyCorpusXL dataset (Section 3) and comprises the following subtasks:
  - **Textual**: participants can only use the provided textual content of the tweets from PolicyCorpusXL for development;
  - **Contextual**: participants can employ for development the textual content of the tweets plus contextual information given to them (i.e., metadata of the tweet and author, friends, retweets, and reply relations).
- **Task B – Cross-domain hate speech detection**: a binary classification task with test data from different domains – i.e., political and religious. The main objective of this task is to explore cross-domain hate speech detection under two evaluation settings:
  - **XPoliticalHate**: the test set consists of tweets from PolicyCorpusXL (as in both the in-domain subtasks above);
  - **XReligiousHate**: the test set consists of tweets from the ReligiousHate corpus (Section 3), for which no development data is provided to participants.

  Moreover, participants are allowed to use any kind of external data (e.g., datasets for other hate domains) and textual and contextual PolicyCorpusXL development data.

3. Dataset and Format

In this section, we describe the dataset creation process (Section 3.1), including data collection, annotation, enrichment, and label distribution. Then, we outline the format used for sharing data to participants (Section 3.2).

3.1. Dataset Creation

We collected data from Twitter after selecting it among existing social media platforms where hatred content could be present. There are two main reasons for this choice. One the one hand, Twitter easily allows the retrieval of a high volume of textual content by using APIs. On the other hand, additional metadata about tweets themselves and their authors can be collected. Furthermore, Twitter users can perform asynchronous actions such as retweeting, replying, and following. This latter aspect allows us to share with HaSpeeDe3 participants not only the text of the tweets and their metadata but also contextual information about the network where the participants of the online debate are situated.

3.1.1. Data Collection

The focus of HaSpeeDe3 is the detection of hate speech in strong polarised debates, in particular concerning political and religious topics. We use two different datasets for the shared task: **PolicyCorpusXL** [24] and **ReligiousHate** [20]. For both of them, tweets have been collected via the Twitter APIs by querying for key terms specific to each topic.

PolicyCorpusXL The dataset contains 7,000 tweets collected employing a snowball sampling from three starting hashtags (#dpcm, #legge, #leggedibilancio). 5,736 tweets have been collected between April and July 2021 and 1,264 between March and May 2020 [22].

ReligiousHate We use the Italian portion of the religious hate speech corpus introduced in [20]. The dataset is composed of 3,000 tweets collected between December

2020 and August 2021 with keywords that refer to the three main monotheistic religions, namely Christianity, Islam and Judaism.

Due to the different nature of the political and religious topics, the protocols used for data collection are not the same; however, in both cases, offensive words have not been used as query terms to minimise biased dataset composition and potential learning shortcuts [25, 26].

3.1.2. Data Annotation

We summarise the annotation procedure followed for PolicyCorpusXL and ReligiousHate below.

PolicyCorpusXL Two Italian experts of communication annotated the entire dataset. The training set has been additionally annotated by a third expert in case of disagreement. 1,000 tweets have been finally discarded in order to artificially augment the portion of hate tweets and provide more information for the classifiers. With this strategy, the number of tweets containing hate increased from 11.8% (a typical percentage obtained with random sampling) to 40.6%.

ReligiousHate Three native speakers of Italian with a background in linguistics and computer science annotated 3,000 tweets about religion that have been collected as described in Section 3. Annotation was performed following a protocol for experts that foresaw in-person discussion rounds and adjudication sessions.

The Inter Annotation Agreement is similar for both the PolicyCorpusXL (Fleiss' $k = 0.53$) and ReligiousHate (Cohen's $k = 0.57$) datasets.

3.1.3. Data Enrichment

Using the Twitter stream APIs we retrieve tweets but we miss their subsequent history in the micro-blog platform. Indeed, since tweets are retrieved at posting time, we are not able to know what happens to them afterwards. In order to follow up the impact of a tweet on the user community after the posting time, we, therefore, use Twitter's APIs also to retrieve information about each tweet a posteriori. This makes it possible to check, for example, the number of times that the tweet has been retweeted or liked over the weeks after its posting time. We also collected a variety of additional information about the author, such as the list of friends and the users that each author retweeted and replied to since about 2018.

3.1.4. Label Distribution

Statistics of the two HaSpeeDe3's datasets are summarised in Table 1. PolicyCorpusXL consists of 7,000 tweets about political debates (5,600 in the development set and 1,400 in the test set), whereas ReligiousHate comprises 3,000 tweets, all belonging to the test set.

Table 1
Statistics of the datasets used for the HaSpeeDe3 shared task. HS: hate speech, ¬HS: non-hate speech.

Split	Dataset	HS	¬HS	Total
dev set	PolicyCorpusXL	3,456	2,144	5,600
	ReligiousHate	—	—	—
test set	PolicyCorpusXL	700	700	1,400
	ReligiousHate	487	2,513	3,000
	Total	4,643	5,357	10,000

3.2. Data Format

The development set consist of 5,600 tweets belonging to PolicyCorpusXL. It is organised in the following files, each including the following fields:

training|test_textual.csv

- anonymized_tweet_id: A pseudo-random integer that identifies the specific tweet and replaces the original tweet id.
- anonymized_text: URLs have been replaced by the placeholder `[URL]` and mentions have been replaced and mapped by a pseudo-random integer that identifies a specific user. For example, given the original content "*@MarioRossi Devi rinascere forse 100 volte per poter solo nominare il tuo incubo #Renzi*", the anonymised tweet is "*203951222958528 Devi rinascere forse 100 volte per poter solo nominare il tuo incubo #Renzi*". This means that the pseudo-random integer 203951222958528 identifies all forms of mentions of @MarioRossi, as the author of a tweet, and as a source or target in friends, retweets, and replies relations from/to @MarioRossi.
- label: `1` for hateful tweets, `0` otherwise. Test set labels have been released through the file test_textual_gold.csv after the competition ended.
- dataset: this field specifies the set (training or test) and whether a tweet belongs to the PolicyCorpusXL or the ReligiousHate dataset. training_textual.csv exclusively contains tweets belonging to the PolicyCorpusXL dataset.

training|test_contextual.csv

- anonymized_tweet_id: A pseudo-random integer that identifies the specific tweet and replaces the original tweet id.
- created_at: The posting date of the tweet.
- retweet_count: The number of times the tweet has been retweeted.
- favorite_count: It indicates approximately how many times this tweet has been liked by Twitter users[1].
- source: The source used for posting the tweet (e.g., Android, iOS, web).
- is_reply: 1 if the tweet is a reply, 0 otherwise.
- is_retweet: 1 if the tweet is a retweet, 0 otherwise.
- is_quote: 1 if the tweet is a quote, 0 otherwise.
- anonymized_user_id: The original author id (if known), replaced by a pseudo-random integer.
- user_created_at: The date when the author created the account.
- statuses_count: The number of tweets posted by the author.
- followers_count: The number of Twitter users that follow the author.
- friends_count: The number of Twitter users that the author follows.
- anonymized_description: The self-description of the author of the tweet. We applied the same anonymisation strategy applied to the field anonymized_text of the file train_textual.csv described above.

The value of some fields could be unavailable or set to 0 if we were unable to recover the metadata of the tweet in 2022 (many months after the posting date), for example, because the tweet has been removed by Twitter, deleted, or made unavailable by the author.

training|test_contextual_friends.csv

- source: A user, identified by anonymized_user_id, that follows the target.
- target: A user, identified by anonymized_user_id, that is followed by the source.

training|test_contextual_retweet|reply.csv

- source: A user, identified by anonymized_user_id, that retweeted target.
- target: A user, identified by anonymized_user_id, that has been retweeted by source.
- date: The day when source retweeted target.
- count: The number of times the source retweeted the target that day.

All sources are authors of at least one tweet in the training corpus, but some authors are missing in this file since it was not possible to recover their friend list.

All files described above are available at the official GitHub page of the task[2].

4. Evaluation Measures

We provide four separate official rankings, one for each subtask. Participants can submit two runs for each subtask. However, participants are not required to participate in all subtasks or to submit 2 runs for each of them.

Systems are evaluated using F_1-score computed over the two binary classes, i.e., hate speech (HS) or non-hate speech (¬HS). Therefore, submissions are ranked by averaged F_1-score over the two classes, according to the following equation:

$$F_{1(avg)} = (F_1^{HS} + F_1^{\neg HS})/2$$

4.1. Baselines

We computed baselines using a simple machine learning model. For Task A - Textual, we employed a Support Vector Classifier trained with a unigram representation of the textual content of the tweet. For Task A - Contextual, we devise a baseline using the same classifier as above, based on a unigram representation of the textual content of the tweet, plus the number of retweets and favourites received by the tweet (retweet_count and favourite_count, see Section 3.2), the author degree computed from the friends network, and the author eigenvector centrality computed from the friends network. A last baseline for both the cross-domain hate speech subtasks employs a Support Vector Classifier with a unigram representation of the textual content of the tweet, trained with the XPoliticalHate and HaSpeeDe2 training sets [27].

In Table 2 we present the results obtained by the baselines on the four subtasks.

5. Task Overview: Participation and Results

A total of six teams participated in the HaSpeeDe3 task. We summarise their contribution below.

[1]Twitter released a number that "indicates approximately how many times th[e] Tweet has been liked by Twitter users": https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/tweet

[2]https://github.com/mirkolai/EVALITA2023-HaSpeeDe3

Table 2
Results obtained by the baselines on the four subtasks.

Task/subtask	$F_{1(avg)}$
Task A - Textual	0.8457
Task A - Contextual	0.8457
Task B - PoliticalHate	0.8458
Task B - ReligiousHate	0.5718

BERTicelli [28] The team submitted results for all the tasks and used all the provided sets of information. They exploited two pre-trained cased LLMs for Italian, namely UmBERTo and Italian BERT. In the pre-processing phase, they turned hashtags into words to reduce noise, they performed fine-tuning and used a 5-fold cross-validation for the Textual subtask, obtaining high scores. For the Contextual subtask, the team adopted an ensemble approach, wherein additional features were added to the fine-tuned models through a GradientBoosterClassifier algorithm. UmBERTo performed competitively in both Textual and Contextual subtasks but the model did not benefit from the addition of contextual features. Italian BERT, on the other hand, performed above the baselines but significantly lower than the task average. Overall, the team performed above the average in the political hate domain and below the average in the religious hate domain.

CHILab [29] The team participated only to the Task A - Textual, i.e., addressing only in-domain political hate speech detection using the provided textual content of the tweets from PolicyCorpusXL for development. They submitted two runs that employ two different models based on BiLSTM. The first one generates embeddings of 768 tokens from AlBERTo and the second one employs fastText for generating 300-dimensional token embeddings. Particular attention was paid to pre-processing. The `[URL]` tag, mention references, and retweet notes were removed since they were not considered relevant. Case sensitivity has been preserved as well as emojis due to the fact that they convey a specific meaning in social media communication in terms of prosody and emotions.

extremITA [30] The team addressed all the tasks using all the provided sets of information made available by the organisers. They also made use of data from all the EVALITA 2023 challenges to build monolithic architectures to tackle all the tasks. Their approaches are based on *i)* the IT5 encoder-decoder model, and *ii)* an instruction-tuned large language model built upon LLaMA. To the goal, for both models, they devised natural language instructions and output templates for each EVALITA task, including HaSpeeDe3. Among their submissions, we observe that the LLaMA-based model achieved better results than the IT5 one on Task A - Contextual, whereas the IT5 model achieved better results on the remainder subtasks.

INGEOTEC The team did not submit a system description report; therefore, we are unable to discuss and analyse their approach. They participated to the Task A - Textual and to the Task B - XReligiousHate considering both the evaluation settings.

LMU [31] The team participated only to the Task B - XReligiousHate considering both the evaluation settings with multitask prompt-training systems. Their systems consist of two steps in which models are *i)* pre-finetuned on external datasets in Italian and English from various domains, *ii)* fine-tuned on the target domain (only applicable to PolicyCorpusXL). As a backbone of their systems, they experimented with both Italian and multilingual pre-trained language models (PLMs). They showed that Italian datasets are more beneficial than the combination of Italian and English ones and that systems based on both Italian and multilingual PLMs achieved similar performance. Their best runs for the political and religious domains are ensembles of prompt-training systems based on Italian and multilingual PLMs.

odang4 [32] The team participated in both Tasks A and B, using only textual information in the former. They based their approach on the assumption that a relation between named entities and abusive language exists. They submitted two different runs. The first one employs enhanced-ALBERTo with triple verbalisation from the Ontology of Dangerous Speech [33] with prompting Davinci model. The second one applies a majority voting criteria among ALBERTo, the enhanced-ALBERTo with triple verbalisation from the Ontology of Dangerous Speech, and the enhanced-ALBERTo with prompting Davinci. For what concerns Task B - XReligiousHate, the multilingual expert-based hate speech/counter-narrative pairs dataset on Islamophobia (CONAN) [34] has been employed too.

5.1. Final Ranking

Table 3 shows the results obtained by the participants for each of the four subtasks. The runs submitted by each participant are highlighted in green. However, when a team submits a run to Task A - Textual, the submission satisfies also Task A - contextual and Task B - XPoliticalHate requirements, therefore it is included in the final ranking. Likewise, when a team submits a run to Task A - contextual, the submission satisfies Task B - XPoliticalHate requirements too. The best results in Task A - Textual, Task A - contextual, and Task B - XPoliticalHate

Table 3
Participants' results for each task and subtask. Results are reported as averaged F_1 scores. Best results are in bold. The runs submitted by each team are highlighted in green, whereas the remaining ones indicate the tasks and subtasks in which the team implicitly participated in.

Team	Task A				Task B			
	textual		contextual		XPoliticalHate		XReligiousHate	
	run 1	run 2	run 1	run 2	run 1	run 2	run 1	run 2
BERTicelli	0.8976	0.8652	0.8976	0.8969	0.8976	0.8969	0.5401	0.5384
CHILab	0.8257	0.8516	0.8257	0.8516	0.8257	0.8516		
extremITA	0.9079	0.9034	0.9079	0.9034	0.9079	0.9034	0.5921	**0.6525**
INGEOTEC	0.8845		0.8845		0.8845		0.5522	
LMU					0.9014	0.8984	0.6458	0.6461
odang4	**0.9128**	0.8950	**0.9128**	0.8950	**0.9128**	0.8950	0.5213	0.4809
avg		0.8826		0.8862		0.8887		0.5744
std		0.0293		0.0288		0.0264		0.0624

are achieved by the **odang4** team with $F_{1(avg)}$ = 0.912, employing the same model without taking advantage of contextual information nor using external data sources. Only **extremITA** and **LMU** (the latter exclusively participated to Task B - XPoliticalHate) reached $F_{1(avg)} >$ 0.9 with at least one of their runs.

extremITA and **LMU** are the only two teams that reached $F_{1(avg)} > 0.6$ in Task B - XReligiousHate. In particular, **extremITA** obtained $F_{1(avg)}$ = 0.6525, with a remarkable improvement with respect to other teams.

All participating systems showed an improvement over the baselines employed for the in-domain political hate speech detection tasks, whereas only two teams outperformed the baseline for Task B - XReligiousHate, proving the complexity of the cross-domain task (Section 5.1).

6. Discussion and Conclusion

Results show that the run #1 submitted by the **odang4** team achieves the best scores across all in-domain tasks. In particular, their approach combining prompting, the Ontology of Dangerous Speech, and the ALBERTo model proved particularly effective in the political domain. However, none of the participants seems to have found a way to effectively exploit contextual information yielding an improvement over textual-only models. This is in line with past studies showing the challenges of embedding contextual information in hate speech detection systems [35].

While the best performance for the in-domain task confirms the state-of-the-art results obtained in similar settings [36], we observe a significant drop in performance (around -0.30 F_1 score on average) for the out-of-domain task. Among the systems, **extremITA** shows a better generalisation capability and yields the best results in this setting. We hypothesise that this happens because their system was built to address all EVALITA challenges, and the only task-specific adaptation is the use of instructions for HaSpeeDe3. Overall, out-of-domain settings still challenge hate speech detection capabilities and still represent a research direction to investigate. Furthermore, approaches that tackle well in-domain hate do not seem to suit the out-of-domain setting, for which different strategies should be pursued.

Acknowledgments

This work has received financial support from the European Union's Horizon Europe research and innovation program under grant agreement No 101070190 (AI4Trust).

References

[1] CENSIS, 50° rapporto sulla situazione sociale del paese 2016, Franco Angeli, 2016.

[2] M. Conover, J. Ratkiewicz, M. Francisco, B. Goncalves, F. Menczer, A. Flammini, Political polarization on Twitter, in: International AAAI Conference on Web and Social Media, ICWSM, Association for the Advancement of Artificial Intelligence, Palo Alto, CA, USA, 2011, pp. 89–96.

[3] H. Watanabe, M. Bouazizi, T. Ohtsuki, Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection, IEEE access 6 (2018) 13825–13835.

[4] F. Celli, E. A. Stepanov, M. Poesio, G. Riccardi, Predicting brexit: Classifying agreement is better than sentiment and pollsters., in: PEOPLES@ COLING, 2016, pp. 110–118.

[5] N. Oliver, B. Lepri, H. Sterly, R. Lambiotte, S. Deletaille, M. De Nadai, E. Letouzé, A. A. Salah, R. Benjamins, C. Cattuto, et al., Mobile phone data for informing public health actions across the covid-19 pandemic life cycle, 2020.

[6] M. Caprolu, A. Sadighian, R. Di Pietro, Characterizing the 2022 russo-ukrainian conflict through the lenses of aspect-based sentiment analysis: Dataset, methodology, and preliminary findings, 2022. URL: https://arxiv.org/abs/2208.04903. doi:10.48550/ARXIV.2208.04903.

[7] E. Elejalde, L. Ferres, E. Herder, The nature of real and perceived bias in chilean media, in: Proceedings of the 28th ACM Conference on Hypertext and Social Media, HT, Association for Computing Machinery, New York, NY, USA, 2017, pp. 95–104. URL: http://doi.acm.org/10.1145/3078714.3078724. doi:10.1145/3078714.3078724.

[8] Y. Theocharis, W. Lowe, Does Facebook increase political participation? Evidence from a field experiment, Information, Communication & Society 19 (2016) 1465–1486.

[9] O. Ștefăniță, D.-M. Buf, Hate speech in social media and its effects on the lgbt community: A review of the current research, Romanian Journal of Communication and Public Relations 23 (2021).

[10] E. Fersini, D. Nozza, P. Rosso, Overview of the evalita 2018 task on automatic misogyny identification (ami), in: Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop, EVALITA 2018, volume 2263, CEUR, 2018.

[11] F. Poletto, M. Stranisci, M. Sanguinetti, V. Patti, C. Bosco, Hate speech annotation: Analysis of an italian twitter corpus, in: 4th Italian Conference on Computational Linguistics, CLiC-it 2017, volume 2006, CEUR-WS, 2017, pp. 1–6.

[12] P. Badjatiya, S. Gupta, M. Gupta, V. Varma, Deep learning for hate speech detection in tweets, in: Proceedings of the 26th International Conference on World Wide Web Companion, 2017, pp. 759–760.

[13] F. Celli, G. Riccardi, A. Ghosh, Corea: Italian news corpus with emotions and agreement., in: Proceedings of CLIC-it 2014, 2014, pp. 98–102.

[14] M. Lai, M. Tambuscio, V. Patti, G. Ruffo, P. Rosso, Stance polarity in political debates: A diachronic perspective of network homophily and conversations on twitter, Data & Knowledge Engineering 124 (2019) 101738. URL: https://www.sciencedirect.com/science/article/pii/S0169023X19300187. doi:https://doi.org/10.1016/j.datak.2019.101738.

[15] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti, Resources and benchmark corpora for hate speech detection: a systematic review, Language Resources & Evaluation 55 (2021) 477–523.

[16] P. Saha, B. Mathew, P. Goyal, A. Mukherjee, Hateminers: detecting hate speech against women, arXiv preprint arXiv:1812.06700 (2018).

[17] E. W. Pamungkas, V. Basile, V. Patti, Misogyny detection in twitter: a multilingual and cross-domain study, Inf. Process. Manag. 57 (2020) 102360. URL: https://doi.org/10.1016/j.ipm.2020.102360. doi:10.1016/j.ipm.2020.102360.

[18] I. Guellil, A. Adeel, F. Azouaou, S. Chennoufi, H. Maafi, T. Hamitouche, Detecting hate speech against politicians in arabic community on social media, International Journal of Web Information Systems (2020).

[19] S. Jaki, T. De Smedt, Right-wing german hate speech on twitter: Analysis and automatic detection, arXiv preprint arXiv:1910.07518 (2019).

[20] A. Ramponi, B. Testa, S. Tonelli, E. Jezek, Addressing religious hate online: from taxonomy creation to automated detection, PeerJ Computer Science 8 (2022) e1128. URL: https://doi.org/10.7717/peerj-cs.1128. doi:https://doi.org/10.7717/peerj-cs.1128.

[21] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[22] A. Duzha, C. Casadei, M. Tosi, F. Celli, Hate versus politics: detection of hate against policy makers in italian tweets, SN Social Sciences 1 (2021) 1–15.

[23] F. Giglietto, N. Righetti, G. Marino, L. Rossi, Multiparty media partisanship attention score. estimating partisan attention of news media sources using twitter data in the lead-up to 2018 italian election, Comunicazione politica 20 (2019) 85–108.

[24] F. Celli, M. Lai, A. Duzha, C. Bosco, V. Patti, Policycorpus xl: An italian corpus for the detection of hate speech against politics, in: Proceedings of the Eighth Italian Conference on Computational Linguistics (CLiC-it 2021), volume 3033 of *CEUR Workshop Proceedings*, CEUR-WS.org, Aachen, Germany, 2022. URL: http://ceur-ws.org/Vol-3033/paper38.pdf.

[25] M. Wiegand, J. Ruppenhofer, T. Kleinbauer, Detection of Abusive Language: the Problem of Biased Datasets, in: 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 602–608. URL:

https://aclanthology.org/N19-1060. doi:10.18653/v1/N19-1060.

[26] A. Ramponi, S. Tonelli, Features or spurious artifacts? data-centric baselines for fair and robust hate speech detection, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 3027–3040. URL: https://aclanthology.org/2022.naacl-main.221. doi:10.18653/v1/2022.naacl-main.221.

[27] M. Sanguinetti, G. Comandini, E. Di Nuovo, S. Frenda, M. Stranisci, C. Bosco, T. Caselli, V. Patti, I. Russo, Overview of the evalita 2020 second hate speech detection task (haspeede 2), in: V. Basile, D. Croce, M. Di Maro, L. C. Passaro (Eds.), Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020), CEUR.org, Online, 2020.

[28] L. Grotti, P. Quick, Berticelli at haspeede3: Fine-tuning and cross-validating large language models for hate speech detection, EVALITA 2023 Eigth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (2023) –.

[29] I. Siragusa, R. Pirrone, Chilab at evalita 2023: Overview of the taks a textual, EVALITA 2023 Eigth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (2023) –.

[30] C. D. Hromei, D. Croce, V. Basile, R. Basili, Extremita at evalita 2023: Multi-task sustainable scaling to large language models at its extreme, EVALITA 2023 Eigth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (2023) –.

[31] V. Hangya, A. Fraserl, Lmu at haspeede3: Multi-dataset training for cross-domain hate speech detection, EVALITA 2023 Eigth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (2023) –.

[32] C. Di Bonaventura, A. Muti, M. A. Stranisci, B. McGillivray, A. Meroño-Peñuela, O-dang4 at hodi and haspeede3: A knowledge-enhanced approach to homotransphobia and hate speech detection in italian, EVALITA 2023 Eigth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (2023) –.

[33] M. A. Stranisci, S. Frenda, M. Lai, O. Araque, A. T. Cignarella, V. Basile, C. Bosco, V. Patti, O-dang! the ontology of dangerous speech messages, in: Proceedings of the 2nd Workshop on Sentiment Analysis and Linguistic Linked Data, European Language Resources Association, Marseille, France, 2022, pp. 2–8. URL: https://aclanthology.org/2022.salld-1.2.

[34] Y.-L. Chung, E. Kuzmenko, S. S. Tekiroglu, M. Guerini, CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 2819–2829. URL: https://aclanthology.org/P19-1271. doi:10.18653/v1/P19-1271.

[35] S. Menini, A. P. Aprosio, S. Tonelli, Abuse is contextual, what about nlp? the role of context in abusive language annotation and detection, arXiv preprint arXiv:2103.14916 (2021).

[36] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, c. Çöltekin, SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020), in: Proceedings of SemEval, 2020.

CHILab at HaSpeeDe3: Overview of the Taks A Textual

Irene Siragusa[1], Roberto Pirrone[1]

[1]Dipartimento di Ingegneria @ Università degli Studi di Palermo, Viale delle Scienze, Edificio 6 90129 - Palermo

Abstract

This technical report illustrates the system developed by the CHILab team for the competition HaSpeeDe3 as part of the EVALITA 2023 campaign. The key idea for HaSpeeDe3 task A - Political Hate Speech Detection - Textual, was to develop different systems arranged as suitable combinations of the Pre-Trained Language Model (PTLM) used for embedding extraction, neural architectures for further elaborations over the embeddings and a classifier. In particular, dense layers, LSTM, BiLSTM and Transformers were used. The best performing system across the ones investigated in this report was made by embeddings extracted via XLM-RoBERTa coupled with BiLSTM that reaches a macro-F1 score of 0.876.

Keywords

hate speech detection, BiLSTM, language model

1. Introduction

The continuous spread and usage of social media has become a problem when dealing with hate online. All social platforms use artificial intelligence techniques to detect and report or remove some dangerous contents in terms of hate or violence. The interest in this respect is also high in the scientific community, in fact different international campaigns for detecting hateful speeches have been proposed in recent years: OffensEval [1, 2], HatEval [3], HaHackathon [4]. Detection of hateful content in Italian has been addressed by the HaSpeeDe evaluation competitions [5, 6].

This paper introduces the architecture proposed by the CHILab team for the EVALITA 2023 campaign [7], and in particular as regards the Hate Speech Detection task (HaSpeeDe3 task A - Political Hate Speech Detection, textual) [8]. The general approach relies on encoding the text into suitable word embeddings that are processed via neural architectures like LSTM, BiLSTM or Transformers. Finally, the output classifier detects the presence of hateful content.

We conceived our pipelines as "minimalist" architectures. No generative models [9, 10] where considered in this respect to derive embeddings. Moreover, we decided not to use fine-tuning in our PTLMs to stress the use of light networks to be trained with low computing resources. Finally, we set up a unique approach for all the tasks we have participated in EVALITA 2023.

The paper is arranged as follows: Section 2 reports a description of our systems along with data pre-processing, while results are reported and discussed in Section 3. Concluding remarks are in Section 4.

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT

irene.siragusa02@unipa.it (I. Siragusa); roberto.pirrone@unipa.it (R. Pirrone)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

2. Description of the system

The focus of HaSpeeDe3 was on political and religious hate, where strong polarized opinions can be found. The data set used in this edition for task A is the PolicyCorpus XL [11] that contains 7000 tweets annotated manually, and a presence of hate labels above 40%. The training data set was released for the campaign with a total of 5600 samples: for developing purposes, the given data set was randomly split in a training and validation set, using a 80-20 ratio, resulting in 4480 and 1120 samples respectively.

2.1. Pre-processing

The `[URL]` tag, mention references, and retweet notes were removed since they were not considered meaningful: in particular, mentions are referred to anonymized accounts thus they add no special information. This was done after an analysis on the most cited words and hashtags[1]. As reported in Table 1, the `[URL]` tag is the most frequent one between classes and adds no information just like the anonymized mentions in the form `@unknown`. Overall, no other relevant words appeared that suggest a strong separation between classes. The same considerations can be done for the hashtags as reported in Table 2.

Although there are some hashtags that are hateful (such as *salvinipagliaccio*, *speranzadimettiti* and *governodeipeggiori*), the most frequent ones are just either politicians' or parties' names, and politics related words, that do not express any polarized content. Moreover, since a strong and significant distinction between hateful and non-hateful hashtags can be done, their information has been used as a word inside the tweet, thus keeping the crucial information, while the hashtag symbol was removed.

[1]for this analysis all the words were reported in their lower case form

CEUR-WS.org/Vol-3473/paper23.pdf

Table 1
Word distribution statistics over the dataset divided per label.

All tweets	freq	NH tweets	freq	H tweets	freq
url	1585	url	1007	url	578
governo	455	unknown	297	governo	160
solo	376	governo	295	solo	133
unknown	364	solo	243	797998657209770	132
fare	276	fare	178	salvini	116
fatto	264	fatto	173	fa	112
fa	256	oggi	165	sempre	104
cosa	254	essere	159	cosa	99
salvini	247	cosa	155	fare	98
essere	247	italia	152	poi	96

Table 2
Hashtags distribution statistics over the dataset divided per label.

All tweets	freq	NH tweets	freq	H tweets	freq
salvini	886	salvini	557	salvini	329
m5s	630	m5s	495	salvinisciacallo	251
conte	419	conte	316	salvinipagliaccio	235
draghi	368	legge	284	governodeipeggiori	223
lega	341	governo	255	speranzadimettiti	179
governo	320	draghi	252	speranzavattene	175
legge	309	lega	218	m5s	135
renzi	306	renzi	212	salviniportasfiga	131
salvinisciacallo	255	pd	181	lega	123
pd	251	politica	170	draghi	116

Table 3
Emoji distribution statistics over the dataset divided per label.

All	freq	NH	freq	H	freq
😂	38	🇮🇹	21	💩	32
💩	32	🤔	15	😂	27
🤣	32	🤣	14	🤡	25
🇮🇹	26	🔵	13	🤣	18
🤔	26	😂	11	😡	15
🤡	26	🔴	10	🤔	11
😡	22	‼️	8	🤮	10
🤮	15	😅	8	🤬	10
👇	15	👇	8	🙄	8
🔴	13	😡	7	👇	7

Similar considerations were made for emojis: also in this case, a strong polarization in the use of emojis did not arise, particularly for the ones that are more associated with disgust and hate (Table 3). Since emojis are deeply used in social media communication, they were kept. No further elaboration were made over the tweets: words were not reported to their lower case form, thus allowing a more accurate extraction of embeddings for the case-sensitive PTLMs. As for emojis, uppercase texts has a specific meaning in social media communication in terms of prosodic and emotions interpretation [12, 13].

2.2. Network architectures

Different models were developed that share the same macro structure shown in figure 1. The key idea was to stress, as much as possible, existent neural architectures for sequence processing, that are LSTM [14], BiLSTM and Transformers [15]. Those architecture are used to further process the extracted embeddings.

After pre-processing, the input sentences were padded to *maxLength* + 2 tokens where *maxLength* is the size of the longest sentence, and the remaining two tokens are respectively the `[CLS]` and the `[SEP]` one. Either a pre-trained language model or a static context-free embedding model were used for embedding generation. In the last case, *fastText* [16] was used that generates a 300 tokens embedding, while a 768 tokens embedding is obtained as usual by the different PTLMs. We used the following Encoder-based Language Models in the experiments: BERT base multilingual cased [17], BERT base italian uncased [18], XLM-RoBERTa [19] and AlBERTo [20] provided by the HuggingFace Transformers library[2]. The embeddings were extracted from the last layer of the PTLMs without fine-tuning. Fine-tuning in these configuration is an option that is not taken into account since the main idea is to stress the use of light

[2]https://huggingface.co/docs/transformers/index

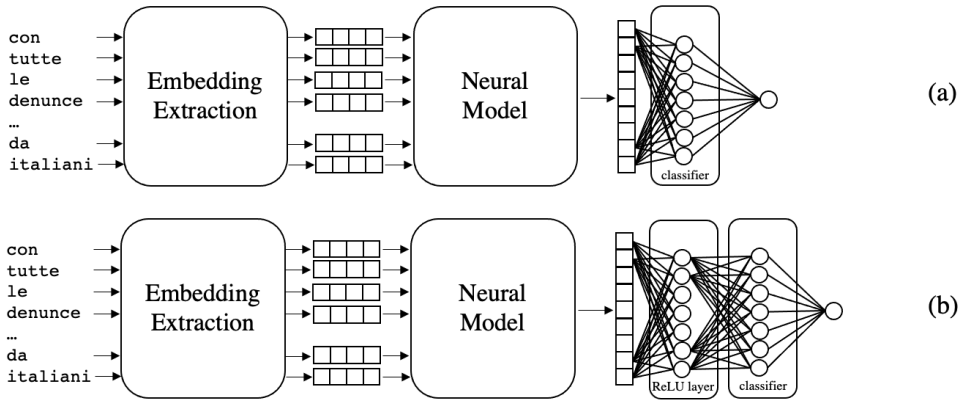


Figure 1: The first proposed architecture (a) has a module for embeddings extraction, a neural module for further processing on the extracted embeddings and a classifier. The second one (b) adds an additional ReLU dense layer.

networks to be trained with low computing resources.

The generated embedding is fed into a module for feature extraction that consists of a LSTM or a BiLSTM or a Transformer[3]. The output feature vector has the same size of the word embedding with the exception of the BiLSTM that generates a double-length output. Finally, the feature vector is passed to a classifier made by either 300 or 768 linear units, depending on the length of the embedding, and a sigmoidal output to achieve binary classification. Some experiments were run by inserting a ReLU dense layer before the aforementioned one with exactly the same size. Those architectures are referred as LSTM-Deep, BiLSTM-Deep and Trasformer-Deep (Figure 1.b).

The illustrated architectures were trained only on the given data set using a machine equipped with two Intel Xeon E5 CPUs 96GB RAM and an NVIDIA TITAN Xp GPU 12GB RAM. Hyperparameters were selected as follows: dropout values in $\{0.1, 0.2\}$, batch size 32, Adam optimizer [21] with learning rate 0.01, and a Binary Cross Entropy loss. Models were trained for a maximum of 1000 epochs with a patience value of 50.

Different feature extractors were implemented using 1, 2 or 3 LSTM/BiLSTM/Transformer layers, but the best results were obtained by the single layer feature extraction modules. In addition the developed models are relative small, where the trainable parameters range from 1M to 10M.

3. Results

The best F1-macro performances obtained on the test set from our models are reported in Table 4. The submitted modes were the best runs with respect to the validation set, namely AlBERTo/BiLSTM (run 1) and fastText/BiLSTM (run 2). After the release of the golden labels, it was possible to measure the actual performance of all the developed systems and this shows up that the XLM-RoBERTa/BiLSTM architecture gives the best results, ranking at the 7th place on the leaderboard, while the submitted runs are at last places as shown in Table 5.

Best results are obtained either when using a PTLM coupled with a LSTM/BiLSTM feature extractor and a single dense layer[4], while the Transformer based networks exploit better a context-free embedding by using a two layer classifier.

It is worth noticing that only the models that use *fastText* benefit from removing stopwords, while the PTLMs perform almost equally over LSTM and BiLSTM configurations as it was expected. In the training phase, AlBERTo outperformed the other PTLMs since it uses a more accurate tokenization compared to the others, and it takes advantage from its inner knowledge: AlBERTo was trained on a corpus of Italian tweets that share the same linguistic macro-structure of the PolicyCorpus. On the other hand, the best model is the one based on XLM-RoBERTa: this can be caused from its tokenizer that owns an inner representation for emojis, and consider them as unique tokens and not as `[UKN]`.

[3]The corresponding architectures are named according the specific neural module

[4]In table 4 some experiments and configurations are not reported, like the BiLSTM-Deep one, because they ran bad with respect to the submitted architectures.

Table 4
The table collects all the relevant results accross the developed architectures. Starred results are the ones submitted to the competition (run1 is BiLSTM model with AlBERTo and run2 is BiLSTM model with fastText), and the bold one is the highest score computed after the golden labels were released. Finally, the underlined results were obtained by removing stopwords from the data. XLM-RoBERTa, fastText and mBERT generate case-sentive embeddings.

	LSTM	LSTM-Deep	BiLSTM	Transformer-Deep
XLM RoBERTa	0.840	0.821	**0.876**	0.836
fastText	0.790	0.726	0.852*	0.861
AlBERTo	0.864	0.834	0.826*	0.852
mBERT	0.857	0.827	0.859	0.831
BERT-it	0.850	0.817	0.868	0.801

Table 5
The table collects the macro F1 results over the test set of the submitted models and the actual best developed model (the starred one). Result of the baseline model is also reported, along with the ranking and expected ranking position.

Run name	Macro F1	Rank
CHILab3*	0.876	7*
CHILab2	0.852	8
Baseline	0.846	9
CHILab1	0.826	10

3.1. Error analysis

Besides the aforementioned differences between the PTLMs used, another analysis was made on the mis-classified tweets by comparing the results of the best architectures (AlBERTo/LSTM, AlBERTo/LSTM-Deep, XLM-RoBERTa and fastText/Transformer-Deep) and the submitted models. Models agree in mis-classifing 32 tweets, and 25 of them are labeled as hateful.

None of these mis-classified tweets contain emoji, that is their presence or absence is not source of bias in those models. Moreover, the majority of those tweets contains hashtags or expressions referring to politicians and topics of interest in the political debate, that per se are not hateful. On the contrary, tweets containing the hashtag *speranzadimettiti*, considered hateful as in 2, can be found in non hateful tweets. In those tweets the author disapproves the governmental behaviour of a minister: in this case it cannot be considered as hateful since it express a negative opinion without insulting.

On the other side, hateful tweets usually contains profanities and vulgar expressions: hateful tweets that are not correctly classified by the developed models, lack of those expressions or put them in an unconventional way (self-obfuscation or embedded in other words) and this lead to their mis-classification.

4. Conclusion

This paper reported the architectures developed by the CHILab team for HaSpeeDe3 task A promoted at the EVALITA 2023 campaign. Our models show that a relatively small classical pipeline made by embedding extraction plus further neural elaboration can have good performance in hate speech detection without the need of fine-tuning PTLMs, and using few computational resources. The use of such "minimalist" architecture is intended to allow for future development of compact explainable models where explicit linguistic knowledge is injected in the network to improve its performance.

Acknowledgments

This work is supported by the PO FESR 2014-2020 grant n. 086201000543, "SCuSi - Smart Culture in Sicily"

References

[1] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval), in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 75–86. URL: https://aclanthology.org/S19-2010. doi:10.18653/v1/S19-2010.

[2] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, c. Çöltekin, SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020), in: Proceedings of SemEval, 2020.

[3] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, M. Sanguinetti, SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational

Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 54–63. URL: https://aclanthology.org/S19-2007. doi:10.18653/v1/S19-2007.

[4] J. A. Meaney, S. Wilson, L. Chiruzzo, A. Lopez, W. Magdy, SemEval 2021 task 7: HaHackathon, detecting and rating humor and offense, in: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Association for Computational Linguistics, Online, 2021, pp. 105–119. URL: https://aclanthology.org/2021.semeval-1.9. doi:10.18653/v1/2021.semeval-1.9.

[5] C. Bosco, D. Felice, F. Poletto, M. Sanguinetti, T. Maurizio, et al., Overview of the evalita 2018 hate speech detection task, in: Ceur workshop proceedings, volume 2263, CEUR, 2018, pp. 1–9.

[6] M. Sanguinetti, G. Comandini, E. D. Nuovo, S. Frenda, M. A. Stranisci, C. Bosco, T. Caselli, V. Patti, I. Russo, Haspeede 2 @ evalita2020: Overview of the evalita 2020 hate speech detection task, EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020 (2020).

[7] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[8] M. Lai, F. Celli, A. Ramponi, S. Tonelli, C. Bosco, V. Patti, Haspeede3 at evalita 2023: Overview of the political and religious hate speech detection task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[9] G. Mialon, R. Dessì, M. Lomeli, C. Nalmpantis, R. Pasunuru, R. Raileanu, B. Rozière, T. Schick, J. Dwivedi-Yu, A. Celikyilmaz, et al., Augmented language models: a survey, arXiv preprint arXiv:2302.07842 (2023).

[10] Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, Z. Liu, Z. Wu, D. Zhu, X. Li, N. Qiang, D. Shen, T. Liu, B. Ge, Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models, 2023. URL: http://arxiv.org/abs/2304.01852. doi:10.48550/arXiv.2304.01852, arXiv:2304.01852 [cs].

[11] F. Celli, M. Lai, A. Duzha, C. Bosco, V. Patti, Policycorpus xl: An italian corpus for the detection of hate speech against politics., in: CLiC-it, 2021.

[12] M. Heath, Orthography in social media: Pragmatic and prosodic interpretations of caps lock, Proceedings of the Linguistic Society of America 3 (2018) 55–1–13. URL: https://journals.linguisticsociety.org/proceedings/index.php/PLSA/article/view/4350. doi:10.3765/plsa.v3i1.4350.

[13] S. Chan, A. Fyshe, Social and emotional correlates of capitalization on Twitter, in: Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, Association for Computational Linguistics, New Orleans, Louisiana, USA, 2018, pp. 10–15. URL: https://aclanthology.org/W18-1102. doi:10.18653/v1/W18-1102.

[14] S. Hochreiter, J. Schmidhuber, Long Short-Term Memory, Neural Computation 9 (1997) 1735–1780. URL: https://doi.org/10.1162/neco.1997.9.8.1735. doi:10.1162/neco.1997.9.8.1735. arXiv:https://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf.

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

[16] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, T. Mikolov, Learning word vectors for 157 languages, in: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018), 2018.

[17] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[18] S. Schweter, Italian bert and electra models, 2020. URL: https://doi.org/10.5281/zenodo.4263142. doi:10.5281/zenodo.4263142.

[19] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, CoRR abs/1911.02116 (2019). URL: http://arxiv.org/abs/1911.02116. arXiv:1911.02116.

[20] M. Polignano, P. Basile, M. de Gemmis, G. Semeraro, V. Basile, AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets, in: Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019), volume 2481, CEUR, 2019. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85074851349&partnerID=40&md5=7abed946e06f76b3825ae5e294ffac14.

[21] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

LMU at HaSpeeDe3: Multi-Dataset Training for Cross-Domain Hate Speech Detection

Viktor Hangya[1,2,*], Alexander Fraser[1,2]

[1]Center for Information and Language Processing, LMU Munich
[2]Munich Center for Machine Learning

Abstract

We describe LMU Munich's hate speech detection system for participating in the cross-domain track of the HaSpeeDe3 shared task at EVALITA 2023. The task focuses on the politics and religion domains, having no in-domain training data for the latter. Our submission combines multiple training sets from various domains in a multitask prompt-training system. We experimented with both Italian and English source datasets as well as monolingual Italian and multilingual pre-trained language models. We found that the Italian out-of-domain datasets are the most influential on the performance in the test domains and that combining both monolingual and multilingual language models using an ensemble gives the best results. Our system ranked second in both domains.

Keywords

hate speech detection, multitask learning, prompt-training

1. Introduction

Due to the sheer amount of social media content, manual filtering for hate speech is impossible which makes building high performance and reliable hate speech classifiers important. To promote research in the field various datasets were built [1, 2], and shared tasks were organized [3, 4, 5], where the best performing systems are based on pre-trained language models (PLMs) [6, 7].

The *HaSpeeDe3* shared task [8] is the third iteration of the series on hate speech detection in Italian social media posts (tweets) organized at EVALITA 2023 [9], focusing on strongly polarized debates in political and religious topics. Two subtasks were organized: Task A – *Political Hate Speech Detection* which on top of textual inputs, allows for the use of contextual information, such as metadata of tweets and authors. Task B – *Cross-domain Hate Speech Detection* involves only textual inputs, however the main objective is to explore cross-domain hate speech detection in the politics and religious domains, by allowing the use of external datasets (open track). In contrast to the politics domain where in-domain training data is given, in the religious domain such data is not provided. Our team participated only in Task B.

Cross-domain training is a crucial problem in machine learning, aiming to build high quality models for the target domain by leveraging labeled samples from out-of-domain sources as well [10]. For hate speech detection, [11] experimented with training classifiers using out-of-domain training examples and showed a significant performance drop on the test sets compared to in-domain training. By simply combining multiple datasets of different domains, including the target domain, they achieved only slight improvements. In a similar work, one source- and one target-domain were explored [12], but the authors showed mixed results, i.e., improvements on some domains but decrease on others. Similarly, [13] applied the general domain adaptation technique of [10] and showed improvements when incorporating some out-of-domain datasets into the final model, even though the approach seemed sensitive to the chosen out-of-domain dataset. In addition, [14] showed negative performance on the target-domain in German by using additional source-domain English training examples.

Following previous work, we rely on transfer learning to leverage out-of-domain (external) datasets to build our classifiers for the political and religious domains. We experiment with various external datasets containing both Italian and English hate speech inputs. Additionally, in contrast to previous work which used datasets with matching label sets, we use corpora annotated with different label sets, e.g., stereotype. To avoid negative results, we combine multiple datasets in a multitask training fashion in order to build robust systems. Additionally, we train our systems in a two-step process, where we first pre-finetune our models on the external datasets, followed by fine-tuning them to the target task. As the basis of our models, we take various PLMs based on the BERT [15] and RoBERTa [16] architectures, including

The Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), September 07–08, 2023, Parma, Italy

*Corresponding author.
hangyav@cis.lmu.de (V. Hangya); fraser@cis.lmu.de (A. Fraser)
https://www.cis.uni-muenchen.de/~hangyav (V. Hangya); https://www.cis.uni-muenchen.de/~fraser (A. Fraser)
0000-0002-5144-3069 (V. Hangya); 0000-0003-4891-682X (A. Fraser)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper24.pdf

both Italian only and multilingual models. Furthermore, in order to facilitate information sharing across the used datasets, we perform prompt-training which eliminates dedicated classification heads for each dataset.

Our experiments show that using only Italian external datasets is more beneficial compared to leveraging English as well. In contrast, we find that both monolingual and multilingual PLMs perform comparably well, and that they can support each other when combining them using model ensembling.[1]

2. Approach

Our approach consists of two steps where we first pre-finetune a given PLM on external datasets (see Section 3.1), followed by in-domain fine-tuning in case of the political domain where such data is provided. Instead of classification heads, we leverage model prompting.

Prompt-Training Prompt-training was shown to be effective and more reliable for various NLP tasks, including classification [17]. Instead of using classification heads on top of PLMs which add additional parameters to the model, it relies on the masked language modeling task (MLM). Using pattern-verbalizer-pairs (PVPs), an input sentence is first transformed using the pattern, e.g., *I hate you.* $\rightarrow$ *Is this hate speech? I hate you. [MASK]*, and the task is to predict the masked token. Finally, the verbalizer maps the highest probability token, out of a set of valid tokens, to labels of a given dataset, e.g., *Yes* $\rightarrow$ HATE or *No* $\rightarrow$ NONHATE. During training all model parameters are fine-tuned using the MLM objective.

Step 1 Given a set of external training corpora ($D_E = \{D_{e_i} : i = 1..N\}$), we randomly select a single dataset D_{e_i} and a batch of samples from it in each training step. For each dataset we apply a dedicated PVP (see Section 3.2) in order to handle datasets of different label sets, and use cross-entropy loss to perform a single model update. This way we mix the available external datasets during pre-finetuning instead of performing a sequential model update which could lead to catastrophic forgetting. Additionally, we make sure that we exhaust all datasets in D_E in each epoch, i.e., the model is trained on each input sample once per epoch.

Step 2 In case of the political test domain, we apply a second round of model fine-tuning given the in-domain training dataset. We follow the same training procedure as in Step 1 but using only a single training corpus instead of multiple corpora. The goal of this step is to specialize our model to the target domain, given the pre-finetuned base model which is already aware of general hate speech language phenomena. This is in contrast to standard multitask training where i) the goal is to build a single model supporting multiple target tasks (datasets) and more importantly ii) which is trained by optimizing a joint objective function across all datasets.

No in-domain training data was provided for the religious test datasets. In this case, we omit Step 2 and apply our model resulted from Step 1 in a zero-shot transfer learning fashion, i.e., the model is only trained on the external (source) datasets but not on the target corpus.

Ensembling To further improve the robustness of our final models, we employ model ensembling to combine the output of multiple models. We ensemble models in two dimensions: we combine models of the same setup but using 3 different random seeds, and models based on different PLM architectures as defined below. We simply take the mean of the probabilities of the considered models for a given input sample.

3. Experiments

3.1. Datasets

Next, we list our external dataset setups followed by the introduction of the official shared task data. We define the following groups of external datasets:

HASPEEDE We leverage Italian datasets from previous HaSpeeDe iterations. More precisely, we take i) the training data containing 2 400 Facebook posts annotated with binary hate speech labels from HaSpeeDe1 [18], ii) 5 470 binary hate speech annotated Twitter posts from HaSpeeDe2 [5] and iii) the same Twitter posts but annotated for binary stereotype detection.

IT Additionally, to the datasets mentioned in the HASPEEDE set, we used further Italian abusive language related datasets. Tweets from the AMI18 misogyny detection shared task [19]: i) 3 200 binary and ii) 1 460 fine-grained (discredit, stereotype, dominance, harassment, derailing) training sets as well as iii) 1 454 binary target detection set (individual, group). Furthermore, we took binary iv) hate (3 271) and v) stereotype (441) annotated training sets from the IHSC corpus [20] containing tweets related to immigrants.

MIXED Finally, to test the effect of leveraging English training data as well, in addition to the datasets contained in the HASPEEDE set we used 7 078 politics related tweets annotated for binary hate speech detection released in [21].

[1] Our code is available at https://cistern.cis.lmu.de/multi_hs

Table 1
PVPs. For each pattern the input sentence in depicted as **X**. For the verbalizers the left-hand side of → indicates the predicted tokens by the PLMs which are assigned to the label on the right-hand side. We note that we used English to Italian machine translation to build the Italian PVPs.

	Pattern
$p1$	**X** → Questo è un discorso di odio? **X** [MASK]
$p2$	**X** → Is this hate speech? **X** [MASK]
$p3$	**X** → Questo è stereotipato? **X** [MASK]
$p4$	**X** → **X** Era [MASK]
$p5$	**X** → **X** Era preso di mira [MASK]
	Verbalizer
$v1$	*Sì* → HATE; *No*→ NONHATE
$v2$	*Yes* → HATE; *No*→ NONHATE
$v3$	*stereotipato* → STEREOTYPE; *predominante* → DOMINANCE; *deragiante* → DERAILING; *molesto* → SEXUAL_HARASSMENT; *screditante* → DISCREDIT
$v4$	*individuale* → INDIVIDUAL; *gruppo* → GROUP

Official HaSpeeDe3 datasets The HaSpeeDe3 shared task focuses on strongly polarized debates in two domains. For the politics domain, the binary hate speech labeled PolicyCorpusXL was made [22], containing 5 600 train and 1 400 test tweets. In the religious domain, the ReligiousHate [23] corpus contains 3 000 test tweets and no training set.

3.2. Setup

PVPs We aimed at keeping our used patterns and verbalizers simple and uniform across datasets. Both patterns and verbs are presented in Table 1. For binary hate and misogyny datasets we used patterns $p1$ and $p2$ for the Italian and English datasets respectively. Similarly, we used $p3$ for the binary stereotype datasets. As verbalizers, we used $v1$ and $v2$ for the two languages. For the AMI18 misogyny fine-grained and target sets we used patterns number $p4$ and $p5$ respectively, with verbalizers $v3$ and $v4$.

Models As the base PLMs we experiment with two monolingual Italian and two multilingual models. AlBERTo was trained purely on Italian social media texts (Twitter in particular), based on the BERT base architecture [24]. We selected this model since it performs well on social media texts. Similarly, we experiment with UmBERTo [25] which is based on the RoBERTa base architecture, and was trained with whole word masking on Italian CommonCrawl corpus. As for the multilingual models, we used the highly popular mBERT [15] and XLM-R [16] PLMs.

We used the OpenPrompt toolkit for implementation [26], and used standard hyperparameter values. Due to memory limitation of Nvidia GTX 1080 Ti however, we used batch size 4 with gradient accumulation steps 4 for BERT based models, while we used batch size 1 with gradient accumulation steps 16 for RoBERTa based models. We train our models for a single epoch in Step 1 of our approach, while we perform early stopping in Step 2 based on the performance on the development set. During the development of our system, we split the official political training set to train/dev/test splits. Since no labeled sets were provided for the religious domain for development, we simulated zero-shot transfer experiments on the politics domain.

Preprocessing We also experimented with two sets of data manipulation methods. To clean tweets, we applied standard Twitter preprocessing steps: user mention and hashtag removal, HTML and repeated character unification. Since hate speech datasets often suffer from label imbalance, we tested random oversampling, class weighting and focal loss. However, none of these approaches led to consistent improvements, thus we omitted these steps from our final systems.

4. Results

We evaluate our systems using macro averaged F_1 scores as it is the official score used in the shared task. First, we present the comparison of various external dataset setups (Table 2), followed by the comparison of different PLMs and their combination with ensembling (Table 3). Finally, we present our official results in Table 4.

External datasets As the baseline system to measure the effectiveness of the external datasets, we only perform Step 2 of our approach, i.e., we fine-tune the off-the-shelf PLM (mBERT) using only the HaSpeeDe3 politics training corpus without any pre-finetuning steps on the external datasets. As mentioned, no in-domain data is provided for the religious domain, thus we perform zero-shot transfer learning, i.e., we only perform pre-finetuning on the external datasets. Additionally, since not even a development set was provided for this domain, we simulate zero-shot transfer on the politics dataset. The gold labels of the religious test set were released after the shared task deadline, thus we are able present (oracle) results for comparison. The results in Table 2 show the positive impact of the external datasets, as the baseline systems were outperformed by a large margin. Comparing the different external dataset setups, we found that they perform comparably. On the politics domain the HASPEEDE setup performed the best, although both IT and MIXED lagged behind with less than half a percentage point in the two-step setting, while on the simulated zero-shot experiments the gap between HASPEEDE and

Table 2
Macro F_1 scores (%) comparing different external dataset setups using mBERT as the base PLM. The baseline system uses the HaSpeede3 training dataset only, while the HaSpeeDe, It and Mixed incorporate the external datasets as well. Pol. depicts the results of our systems for the politics domain, zero Pol. our zero-shot experiments on the politics domain simulating the missing train set of the religious domain and gold Rel. shows results on the gold religious test set after its release. We bold the **best** setup.

	Pol.	zero Pol.	gold Rel.
baseline	84.48	-	52.34
HaSpeeDe	**86.61**	**63.99**	62.02
It	86.43	61.82	61.69
Mixed	86.21	-	**62.36**

Table 3
Macro F_1 scores (%) comparing different monolingual and multilingual PLMs, as well as model ensembles (mono: AlBERTo and UmBERTo; mix: AlBERTo, UmBERTo, mBERT and XLM-R). We use the HaSpeeDe external dataset setup. We highlight the best individual and **ensemble** models.

	Pol.	zero Pol.	gold Rel.
AlBERTo	89.92	63.00	64.16
UmBERTo	88.77	62.16	62.81
mBERT	86.61	63.99	62.02
XLM-R	86.21	55.45	61.98
mono-ens.	**91.44**	**61.57**	64.58
mix-ens.	90.95	60.71	**64.61**

It[2] is around 2 percentage points. These findings indicate that the misogyny detection tasks in the It setup could be slightly detrimental to the binary hate speech detection task. Furthermore, the additional English politics related dataset in the Mixed setup does not lead to further improvements on the politics domain, although they are from the same domain, indicating that leveraging only Italian external datasets is an important factor. Looking at the results on the gold religious test set, we found similar trends. The use of additional training datasets on top of the HaSpeeDe3 politics training set improves the performance[3]. Although the HaSpeeDe set performs well, interestingly the best performance was achieved by Mixed which includes English politics tweets, which needs further investigations. Nonetheless, based on these findings, we used the HaSpeeDe setup in our final system submission and in the following experiments.

[2]Due to the inclusion of politics related training data in the baseline and Mixed setups, these are not applicable in the simulated zero-shot case.

[3]Note that we also included the politics HaSpeeDe3 train set in the HaSpeeDe, It and Mixed sets when training our models for the religious domain.

Table 4
Our final results as reported by the shared task organizers.

	XPoliticalHate	XReligiousHate
Run 1 (mono-ens.)	**90.14**	64.58
Run 2 (mix-ens.)	89.84	**64.61**

Model variations In Table 3 we compare the mentioned 4 PLMs and their combinations. In the mono-ens. ensemble setup we combine the monolingual Italian models (AlBERTo and UmBERTo), while in mix-ens. all PLMs (AlBERTo, UmBERTo, mBERT and XLM-R). We found that the monolingual models outperform multilingual models in most cases, especially on the politics domain. AlBERTo has the best performance on average which is due to its pre-training on social media content. Interestingly, comparing BERT (AlBERTo and mBERT) and RoBERTa (UmBERTo and XLM-R) architectures, the former outperform the latter, which is a somewhat contradictory result as the latter often performs better. The ensemble results, however, show that although the results of different PLMs vary, they can support each other and by ensembling their outputs the performance can be further increased. Similarly, as for the individual models, the monolingual ensemble performed the best during our system development, however the combination of all models does not lag much behind. Furthermore, mix-ens. outperformed mono-ens. on the gold religious test set.

Final Submission The shared task allowed two submitted runs for each domain. Based on our findings during development, our official systems were mono-ens. (Run 1) and mix-ens. (Run 2) using the HaSpeeDe external dataset setup. We note that in the case of the religious domain, we also include the HaSpeeDe3 politics training set as an external dataset. Our official results are shown in Table 4. We achieved the second-best result in both domains.

5. Conclusions

We presented the LMU Munich team's systems at the HaSpeeDe3 shared task, participating in the cross-domain hate speech detection task. Our approach involves a two-step method for the politics domain: pre-finetuning using external datasets followed by a second step of fine-tuning on the target domain. In case of the religious domain, we used a zero-shot transfer setup involving training on the external datasets only. Additionally, we performed prompt-training instead of the use of classification heads in order for a more seamless combination of external datasets of different label sets. By comparing various external datasets, including both Italian and English, we found that Italian datasets are

more beneficial. Similarly, by comparing various PLMs we found that individually monolingual models perform better than multilingual models. On the other hand, combining multiple PLMs with model ensemble, we found that different models can support each other leading to improved performance. Our best result on the political domain was achieved by combining monolingual PLMs only, while combining all PLMs performed the best on the religious domain.

Acknowledgments

We thank the anonymous reviewers for their helpful feedback and the Cambridge LMU Strategic Partnership for funding for this project.[4] The work was additionally supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (No. 640550) and by the German Research Foundation (DFG; grant FR 2829/4-1).

References

[1] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on Twitter, in: Proceedings of the NAACL Student Research Workshop, 2016, pp. 88–93. URL: https://aclanthology.org/N16-2013.

[2] O. de Gibert, N. Perez, A. García-Pablos, M. Cuadros, Hate speech dataset from a white supremacy forum, in: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), 2018, pp. 11–20. URL: https://aclanthology.org/W18-5102.

[3] T. Mandl, S. Modha, P. Majumder, D. Patel, M. o. Dave, C. Mandlia, A. Patel, Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in Indo-European languages, in: 11th meeting of the Forum for Information Retrieval Evaluation, 2019, pp. 14–17. URL: http://ceur-ws.org/Vol-2517/T3-1.pdf.

[4] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, M. Sanguinetti, SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter, in: Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 54–63. URL: https://aclanthology.org/S19-2007.

[5] M. Sanguinetti, C. Gloria, E. Di Nuovo, S. Frenda, M. A. Stranisci, C. Bosco, C. Tommaso, V. Patti, R. Irene, et al., Haspeede2@ evalita2020: Overview of the evalita 2020 hate speech detection task, in: Proc. EVALITA, 2020, pp. 1–9.

[6] E. Lavergne, R. Saini, G. Kovács, K. Murphy, Thenorth@ haspeede 2: Bert-based language model fine-tuning for italian hate speech detection, in: Proc. EVALITA, volume 2765, 2020.

[7] T. Tran, Y. Hu, C. Hu, K. Yen, F. Tan, K. Lee, S. R. Park, HABERTOR: An efficient and effective deep hatespeech detector, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 7486–7502. URL: https://aclanthology.org/2020.emnlp-main.606.

[8] M. Lai, F. Celli, A. Ramponi, S. Tonelli, C. Bosco, V. Patti, Haspeede3 at evalita 2023: Overview of the political and religious hate speech detection task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), 2023.

[9] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), 2023.

[10] H. Daumé, Frustratingly Easy Domain Adaptation, in: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, 2007, pp. 256–263. URL: https://www.aclweb.org/anthology/P07-1033.

[11] G. Glavaš, M. Karan, I. Vulić, XHate-999: Analyzing and Detecting Abusive Language Across Domains and Languages, in: Proceedings of the 28th International Conference on Computational Linguistics, 2021, pp. 6350–6365. URL: https://www.aclweb.org/anthology/2020.coling-main.559/.

[12] M. A. Rizoiu, T. Wang, G. Ferraro, H. Suominen, Transfer learning for hate speech detection in social media, arXiv:1906.03829 (2019). URL: https://arxiv.org/pdf/1906.03829.pdf.

[13] M. Karan, J. Šnajder, Cross-Domain Detection of Abusive Language Online, in: Proceedings of the 2nd Workshop on Abusive Language Online, 2018, pp. 132–137. URL: https://www.aclweb.org/anthology/W18-5117.

[14] M. Wiegand, A. Amann, T. Anikina, A. Azoidou, A. Borisenkov, K. Kolmorgen, I. Kröger, C. Schäfer, Saarland University's Participation in the GermEval Task 2018 (UdSW) – Examining Different Types of Classifiers and Features, in: Conference on Natural Language Processing (KONVENS 2018), 2018, pp. 21–26. URL: https://www.oeaw.ac.at/fileadmin/subsites/academiaecorpora/PDF/GermEval2018_Proceedings.pdf#page=27.

[15] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for

[4] https://www.cambridge.uni-muenchen.de

Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 4171–4186. URL: https://www.aclweb.org/anthology/N19-1423.pdf.

[16] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised Cross-lingual Representation Learning at Scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440–8451. URL: https://www.aclweb.org/anthology/2020.acl-main.747/.

[17] T. Schick, H. Schütze, Exploiting cloze-questions for few-shot text classification and natural language inference, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 2021, pp. 255–269. URL: https://aclanthology.org/2021.eacl-main.20.

[18] C. Bosco, D. Felice, F. Poletto, M. Sanguinetti, T. Maurizio, et al., Overview of the evalita 2018 hate speech detection task, in: Ceur workshop proceedings, 2018, pp. 1–9.

[19] E. Fersini, D. Nozza, P. Rosso, Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI), Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (2018) 59–66. URL: https://pdfs.semanticscholar.org/05d5/17f3fa5f47b16265b378c81a0839ed760ba0.pdf.

[20] M. Sanguinetti, F. Poletto, C. Bosco, V. Patti, M. Stranisci, An Italian Twitter corpus of hate speech against immigrants, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018. URL: https://aclanthology.org/L18-1443.

[21] C. Toraman, F. Şahinuç, E. Yilmaz, Large-scale hate speech detection with cross-domain transfer, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 2215–2225. URL: https://aclanthology.org/2022.lrec-1.238.

[22] F. Celli, M. Lai, A. Duzha, C. Bosco, V. Patti, Policycorpus XL: An Italian Corpus for the Detection of Hate Speech Against Politics., in: In Proceedings of the Eighth Italian Conference on Computational Linguistics, 2021. URL: https://ceur-ws.org/Vol-3033/paper38.pdf.

[23] A. Ramponi, B. Testa, S. Tonelli, E. Jezek, Addressing religious hate online: from taxonomy creation to automated detection, PeerJ Computer Science 8 (2022) e1128. URL: https://peerj.com/articles/cs-1128.

[24] M. Polignano, V. Basile, P. Basile, M. de Gemmis, G. Semeraro, Alberto: Modeling italian social media language with bert, IJCoL. Italian Journal of Computational Linguistics 5 (2019) 11–31. URL: https://journals.openedition.org/ijcol/472.

[25] L. Parisi, S. Francia, P. Magnani, Umberto: an italian language model trained with whole word masking, https://github.com/musixmatchresearch/umberto, 2020.

[26] N. Ding, S. Hu, W. Zhao, Y. Chen, Z. Liu, H. Zheng, M. Sun, OpenPrompt: An open-source framework for prompt-learning, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2022, pp. 105–113. URL: https://aclanthology.org/2022.acl-demo.10. doi:10.18653/v1/2022.acl-demo.10.

BERTicelli at HaSpeeDe 3: Fine-tuning and Cross-validating Large Language Models for Hate Speech Detection

Leonardo Grotti[1,2], Patrick Quick[1]

[1]Universiteit Antwerpen, Faculty of Arts, Prinsstraat 13, B-2000, Antwerp
[2]CLiPS Research Center, University of Antwerp, Belgium

Abstract
The present paper describes the results from the experiments carried out for the HaSpeeDe 3 shared task, an Italian-language Hate Speech (HS) detection task, at EVALITA 2023. Two BERT-based language models were selected: UmBERTo (cased) and Italian BERT (cased). For the Textual task, the models were fine-tuned and cross-validated across 5 folds. For the Contextual task, we adopted an ensemble approach: the additional features were added to the fine-tuned models through the GradientBoosterClassifier algorithm. The models perform better than the baselines (DummyClassifier and LogisticRegression) and above the average performance of participants in the shared task. While the addition of contextual features did not improve the performance of UmBERTo, it significantly bettered the results obtained with Italian BERT.

Keywords
Hate Speech detection, Italian language, BERT-based language models, Fine-tuning, Contextual features

1. Introduction

The escalating issue of toxic language has been amplified by the rapid growth in social media usage over the past decade [1]. Platforms such as Facebook and Twitter have transformed the way individuals interact, making it faster and often anonymous, thereby creating an ideal environment for the propagation of harmful content [2]. Furthermore, previous studies have shown that this content can be targeted at and posted by both individuals and groups, inciting and driving violent acts in the offline world [2, 3, 4].

As such, countering the phenomenon of toxic language has garnered significant attention from legal authorities, social media platforms, and companies [5]. Platforms like Facebook, Twitter, YouTube, and other websites have taken measures to combat toxic language by implementing bans. However, research has pointed out the limitations of companies' control systems and their heavy reliance on user reports to identify problematic comments or posts [6]. The manual filtering of messages containing toxic language has proven to be not only time-consuming but also detrimental to human annotators [7]. Additionally, studies have revealed that human-labeled data can be influenced by individual annotators' biases [8].

Such interest was reflected in the field of Natural Language Processing (NLP), which has witnessed a surge in interest and popularity in automatic toxic language detection [8]. Researchers aim to develop models that can alleviate the harm caused by online HS [9]. Automating the detection process not only overcomes the challenges of manual filtering but also enables efficient analysis of large volumes of data.

As a reminder, we here use the terms HS as an umbrella term and do not distinguish between its subcategories on a theoretical level. For a more extensive discussion of HS hierarchies and definitions, refer to Zampieri [7] and Caselli et al. [10]. It is worth noting that scholars often do not agree on what constitutes HS and how it differs from, e.g., offensive or aggressive language [11].

2. Related Work

As we have mentioned, the growing interest in addressing toxic language is evident in the numerous tasks dedicated to its detection and its various subcategories. These include Aggression Identification [12], Offensive Language Identification [7], and HS detection in Italian Facebook and Twitter messages [13], among others. Over time, the quantity and quality of available models for toxic language detection have significantly increased. Markov, Gevers, and Daelemans (2021) note that the advent of transformer-based pre-trained language models, coupled with the abundance of user-generated content on social media, has greatly improved detection accuracy.

Despite the overall improvement of the models, a series of challenges remain. For instance, it has been shown how the lack of data in languages other than English [14] has exacerbated already existing issues, such as the high occurrence of code words and misspellings in HS text

EVALITA 2023: Final Workshop of the 8th evaluation campaign, September 08–09, 2021, Parma, IT
lgrotti@uantwerpen.be (L. Grotti);
patrick.quick@student.uantwerpen.be (P. Quick)
https://github.com/corvusMidnight (L. Grotti);
https://github.com/patrickquick (P. Quick)
0000-0001-7914-3191 (L. Grotti)
© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper25.pdf

[9]. Furthermore, the often-low agreement between human annotators has also been reported to hinder models' performance [15]. Finally, the biggest challenge to date remains the generalisability of models for HS classification: i.e., a model's ability to perform in cross-domain (i.e., coming from different domains, such as Facebook and Twitter) and cross-genre (i.e., text belonging to different genres, e.g., social media posts and journal articles) tasks [15, 8].

To deal with these issues, scholars have adopted different approaches: machine learning algorithms have been improved through the addition of linguistic features [16, 17], word lists [18], and syntactic features [19]. In recent years, researchers have also experimented with complex ensembles of both neural [17] and non-neural [1] models.

In the latter case, they [17] leveraged the large amount of data on which large language models are trained and the increased availability of sentence and sub-word embeddings. In short, large language models possess numerical representations for both smaller (sub-embeddings) and larger (sentence-embeddings) language units. This enables them to capture not only misspellings and rare linguistic forms but also the overall meaning of sentences. Furthermore, due to their pretraining process, these models already have an understanding of language structure, enhancing their performance in capturing the nuances of a text [9, 15][1].

3. Task and Method

3.1. Data

Data set	¬HS Tweets	HS Tweets
Development	3456	2144
Test (in)	700	700
Test (out)	2513	487

Table 1
Count of Hate Speech labels in task data.

The task organizers [20] provided development data consisting of 5,600 Italian-language tweets from the Policycorpus XL corpus, a manually-annotated HS corpus [21]. The testing data consists of one subset of in-domain data and one subset of out-of-domain data. The in-domain data consists of 1,400 Italian-language tweets from Policycorpus XL, and the out-of-domain data consists of 3,000 Italian-language tweets from the Italian subset of the ReligiousHate corpus, a manually-annotated religious HS corpus [22].

[1]For an extensive explanation of how these factors have improved the performance of HS detection systems, see Yin and Zubiaga 2021

The development data has been anonymized, with a tweet's identifying information and the user's identifying information both mapped to pseudo-random integers. Placeholders have replaced all instances of URLs, and hashtags have been preserved. The data is marginally imbalanced, with a lesser presence of tweets annotated as HS. The accompanying contextual information is comprised of attributes of Tweet objects and User objects [23].

The structure and anonymization of both the in-domain and out-of-domain testing data follows that of the development data. Upon reconciliation with the gold data, we can see that the in-domain subset is balanced, whereas the out-of-domain subset is significantly unbalanced.

3.2. Task

The task proposal for HaSpeeDe 3 focuses on Italian-language HS detection on Twitter and consists of two tracks with two sub-tasks each. Both Task A and B are binary classification problems to determine whether a tweet contains HS or not. The two sub-tasks in Task A are: (i) Textual, where participants can only use the provided textual content of the tweets from Policycorpus XL for development, and (ii) Contextual, where participants can employ the textual content of the tweets along with the accompanying contextual information.

Task B deals with test data from different domains. The evaluation includes two sub-tasks: XPoliticalHate, where the test set consists of tweets from Policycorpus XL, and XReligiousHate, where the task focuses on recognizing religious hate using tweets from the ReligiousHate corpus. Participants are allowed to use external data, including from other hate domains, and are not restricted to the textual and contextual data provided in Policycorpus XL for development.

The final results in the HaSpeeDe 3 task are evaluated and ranked by computing the F1-score over each class to arrive at an averaged F1-score.

3.3. Models

Fine-tuning, a common technique in NLP, is a form of transfer learning that involves training a pre-trained model on new data to adapt it for a specific downstream task [24]. As mentioned in Section 2, there are notable benefits in fine-tuning models when it comes to HS detection tasks. Furthermore, this approach has been widely used in Italian HS detection, see, e.g., Eric et al. (2020), Tamburini (2020), Nozza et al. (2022). Thus, we fine-tune two large pre-trained language models:

- **UmBERTo-commoncrawl-cased** (Run 1) is a RoBERTa-based model using the OSCAR (Open

Super-large Crawled ALMAnaCH coRpus) Italian large corpus. The model is used for both Named Entity Recognition (NER) and Part Of Speech (POS) tagging and reached excellent performance on different datasets.

- **bert-base-italian-cased** (Run 2) is a BERT-based model which was trained on two million tokens and over 13GB of data. The model was pre-trained on a combination of data which includes the OPUS corpus as well as a Wikipedia dump. Note that for ease of readability, we will now refer to this model as BERT-ita.

For the Textual tasks (both in- and out-of-domain) our experimental setup consists of three stages: To begin with, we apply two basic preprocessing steps, which consist of substituting the pseudo-random user identifiers (e.g., '@12020569') with '@USER' and removing the hash symbol (i.e., '#') from hashtags. Such steps are applied to avoid excessively long tweets[2] and remove unnecessary noise.
Then, both models are fine-tuned and cross-validated across five epochs using PyTorch Trainer and the Transformers library. The development data is shuffled and divided into five folds. For each fold, the models are fine-tuned on 80% of the development data and evaluated on the remaining 20% across 10 epochs with an EarlyStopping patience of 3. We employ cross-validation to ensure that the obtained results are not dependent on a particular data split but rather generalize well across multiple folds.

During this stage, we also tune the learning rate ($1^{e\text{-}3}$, $2^{e\text{-}5}$, and $5^{e\text{-}05}$)[3]. We do not tune batch size as the test data was not available at this stage and increasing batch size may have improved development set performance but worsened generalizability on unseen data (see He et al., 2019). Once the stability of the results has been established through cross-validation, the models are fine-tuned on 85% of the training data[4] (after shuffling) and the resulting model is saved and used to output predictions on both test sets.

For the Contextual task, additional features are incorporated into the model using GradientBoostingClassifier. This ensemble algorithm sequentially trains weak models, resulting in a strong model that is a weighted combination of the weak models. Unlike other algorithms, GradientBoostingClassifier employs decision trees as weak learners and is optimized through gradient descent. To do so, the output labels for both BERT-based models together with the additional features are used as input features for GradientBoostingClassifier.

To further assess the performance of our models, we build four baselines: one LogisticRegression and one DummyClassifier for each task. For Textual and Contextual tasks, the models were trained on the textual data and evaluated on the in-domain test data. For the cross-domain task, they were trained on the same textual data but evaluated on the out-of-domain test set instead. It is worth mentioning that no additional data was used at any stage of our experiments.

4. Results and Discussion

In this section, we describe the results obtained for HaSpeeDe 2023. For each model, we report precision, recall, and F1 score (for both classes). We submitted results for every sub-task except for Task B's XpoliticalHate sub-task. All results are compared with the respective baselines.

4.1. Baselines

As a reference point, Table 2 first presents the baseline results for both in- and out-of-domain tasks for each class. The DummyClassifier performs slightly above random chance for the in-domain task, with an average F1 score of 0.52. However, the out-of-domain results are—as expected—poorer, reaching an average F1 score of 0.42.

LogisticRegression, on the other hand, achieves competitive results, with average F1 scores of 0.86 for in-domain data and 0.52 for out-of-domain data. However, upon further inspection, we can observe how LogisticRegression is fairly limited. With a precision of 0.80 for the non-hate speech (¬HS) class, the model exhibits a relatively high rate of false positives. Additionally, the recall of 0.96 for ¬HS implies a high rate of true positive instances captured but at the expense of potentially overlooking some true negatives. These results suggest that the model may be overly biased towards predicting instances as ¬HS, potentially missing some actual HS instances. Similarly, while a high precision of 0.953 is achieved for the HS class, the model showcases a fairly low recall of 0.75. In turn, this pattern implies the model's inability to identify a significant portion of actual HS instances, resulting in false negatives.

It is worth mentioning that the high performance of LogisticRegression in the in-domain task is likely related to the balanced nature of the data (700 HS v. 700 ¬HS). When out-of-domain, unbalanced test data (see Table 1) is used, performance drastically drops.

[2] The presence of multiple user tags in some of the tweets caused a mismatch in Tensor size and consequently a RuntimeError.

[3] These learning rates are found in Nozza et al. (2022), HuggingFace's fine-tuning guide, and in the standard training parameters, respectively.

[4] Such a configuration is selected to mirror the task's original train-test split of 5600-1400, see Celli et al. (2021).

	Class	Precision	Recall	F1
		In-domain		
Dummy	¬HS	0.518	0.504	0.511
	HS	0.517	0.530	0.523
LogReg	¬HS	0.800	0.963	0.874
	HS	0.953	0.759	0.845
		Out-of-domain		
Dummy	¬HS	0.827	0.485	0.612
	HS	0.152	0.474	0.230
LogReg	¬HS	0.844	0.926	0.883
	HS	0.237	0.119	0.158

Table 2
Results for baseline models for in-domain data and out-of-domain data.

4.2. Task A

Our models achieve competitive results in both Task A's sub-tasks[5], as shown in Table 3 (Textual) and Table 4 (Contextual) for each class. Starting from the former, both models perform above both baselines. However, there seems to be a substantial difference between UmBERTo's (Run 1) and BERT-ita's (Run 2) performance: while the first reaches an F1 average of 0.89, the second reaches 0.86, with a difference between the scores of over .03. Indeed, even if the second run's results are close to the LogisticRegression baseline, the model's predictions (i.e., precision and recall) are more balanced across the two classes. Thus, the F1 for the HS class is higher for BERT-ita compared to the baseline.

The reason for the discrepancy between the two models' performance is likely related to the size of the pre-training data: UmBERTo was trained on over 70GB (against the 13GB of BERT-ita). As such, the model likely has more sub-embeddings and sentence-embeddings available, which in turn allows for better results.

For the Contextual sub-task (Table 4), we included a set of extra features (i.e., 'anonymized description', 'retweet count', 'favorite count', 'is reply', 'is retweet', 'is quote', 'statuses count', 'followers count', and 'friends count') to the output labels through GradientBoostingClassifier. Both models once again reach competitive results. While the first run's results are not affected by the inclusion of contextual features, BERT-ita (Run 2) significantly benefits from their addition. The model performs on the same level as UmBERTo, with an F1 of 0.902 for ¬HS and 0.892 for HS. The inclusion of contextual information during the training stage likely enables BERT-ita to capture more diverse linguistic patterns and generalize better to the classification task.

[5]Note that the overall F1 average for each model and sub-task can be found in Table 7 below.

	Class	Precision	Recall	F1
Run 1	¬HS	0.861	0.949	0.903
	HS	0.943	0.847	0.892
Run 2	¬HS	0.824	0.930	0.874
	HS	0.920	0.801	0.856

Table 3
Results for Task A Textual sub-task.

	Class	Precision	Recall	F1
Run 1	¬HS	0.861	0.949	0.903
	HS	0.943	0.847	0.892
Run 2	¬HS	0.860	0.949	0.902
	HS	0.943	0.846	0.892

Table 4
Results for Task A Contextual sub-task and Task B in-domain sub-task.

4.3. Task B

Though we did not formally submit results to Task B's sub-task XPoliticalHate, we met the requirements of the task by submitting results for the Contextual sub-task of Task A, for which the same test data was used. We will thus report results for both sub-tasks of Task B, referring to Table 4 for the results of Task B's sub-task XPoliticalHate.

Our model performs competitively in the XPoliticalHate sub-task, which made use of in-domain test data, while our model for the sub-task XReligiousHate performed poorly in the context of out-of-domain test data. We made no consideration regarding the XPoliticalHate sub-task, as we did not take any additional steps.

The models' performance on out-of-domain data (Table 5) is much lower than the average F1 score (0.57) but still higher than the baseline (0.52). Such low scores may relate to the imbalance between the two classes in the test data and to limitations in transfer learning. As noted by Ada et al. (2019), performance on the source task may not reflect performance on the target task. Also, the model may overfit on the data on which it was fine-tuned [30].

	Class	Precision	Recall	F1
Run 1	¬HS	0.849	0.950	0.897
	HS	0.330	0.127	0.184
Run 2	¬HS	0.848	0.942	0.893
	HS	0.306	0.131	0.184

Table 5
Results for Task B out-of-domain sub-task.

Overall, our models have consistently outperformed the baselines, demonstrating significant improvements across the board. However, it is worth noting that Table 6 reveals that some runs were below the competition's averages. In particular, Run 2 in Task A (Textual) and

Task B (XReligiousHate) failed to meet our expectations, as discussed in detail in Sections 4.2 and 4.3. These underperforming results can be attributed to the previously highlighted factors.

Task	Sub-task	Model	F1 avg
A	Textual	Run 1	0.89759
		Run 2	0.86516
		Avg	0.88263
	Contextual	Run 1	0.89759
		Run 2	0.89687
		Avg	0.88616
B	XPoliticalHate	Run 1	0.89759
		Run 2	0.89687
		Avg	0.88866
	XReligiousHate	Run 1	0.54011
		Run 2	0.53841
		Avg	0.57439

Table 6
F1 averages for our models and the average for all models submitted to the task.

5. Conclusion

In this paper, we introduced two fine-tuning techniques to detect Italian-language HS in Twitter's posts and replies. We were asked to address the issue in two different tasks with two sub-tasks each. Two models were fine-tuned and 5-cross-validated: UmBERTo and BERT-ita. Task A was comprised of a Textual and a Contextual sub-task: here, UmBERTo performed competitively in both sub-tasks, reaching above the baseline and competition average. However, the model did not benefit from the addition of contextual features. BERT-ita, on the other hand, performed above the baselines but significantly lower than the task average. In contrast to UmBERTo, BERT-ita's results improved significantly, reaching the first model's performance.

For Task B, we did not submit any results for the XPoliticalHate sub-task. As such, the results obtained for Task A (Contextual) were assumed to be valid for this sub-task given the test data was the same. Finally, both our models performed well below the competition average for the out-of-domain task.

Future work should look at the potential benefits of including additional training data for the out-of-domain task. Also, the addition of contextual features could be tested in combination with different language models.

References

[1] M. K. Aljero, N. Dimililer, A novel stacked ensemble for hate speech recognition, Applied Sciences 11 (2021) 1–15. URL: http://dx.doi.org/10.3390/app112411684. doi:10.3390/app112411684.

[2] F. Del Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, M. Tesconi, Hate Me, Hate Me Not: Hate Speech Detection on Facebook., ITASEC (2017) 86–95.

[3] A. A. Siege, Online Hate Speech, Cambridge University Press, 2020. URL: https://doi.org/10.1017/9781108890960[.

[4] K. P. De Maiti, D. Fišer, N. Ljubešić, Nonstandard linguistic features of Slovene socially unacceptable discourse on Facebook, Znanstvena založba Filozofske fakultete (2020).

[5] M. Sanguinetti, F. Poletto, C. Bosco, V. Patti, M. Stranisci, An Italian Twitter Corpus of Hate Speech against Immigrants, Language Resources and Evaluation (2018) 1–8.

[6] M. Sanguinetti, G. Comandini, E. D. Nuovo, S. Frenda, M. Stranisci, C. Bosco, T. Caselli, V. Patti, I. Russo, HaSpeeDe 2 @ EVALITA2020: Overview of the EVALITA 2020 Hate Speech Detection Task, EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020 (2020) 93–101. URL: http://dx.doi.org/10.4000/books.aaccademia.6897. doi:10.4000/books.aaccademia.6897.

[7] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval), Proceedings of the 13th International Workshop on Semantic Evaluation (2019). URL: http://dx.doi.org/10.18653/v1/s19-2010. doi:10.18653/v1/s19-2010.

[8] I. Markov, W. Daelemans, Improving Cross-Domain Hate Speech Detection by Reducing the False Positive Rate, Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda (2021). URL: http://dx.doi.org/10.18653/v1/2021.nlp4if-1.3. doi:10.18653/v1/2021.nlp4if-1.3.

[9] W. Yin, A. Zubiaga, Towards generalisable hate speech detection: a review on obstacles and solutions, PeerJ Computer Science 7 (2021) e598. URL: http://dx.doi.org/10.7717/peerj-cs.598. doi:10.7717/peerj-cs.598.

[10] T. Caselli, V. Basile, J. Mitrović, M. Granitzer, HateBERT: Retraining BERT for abusive language detection in English, in: Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021), Association for Computational Linguistics, Online, 2021, pp. 17–25. URL: https://aclanthology.org/2021.woah-1.3. doi:10.18653/v1/2021.woah-1.3.

[11] T. Caselli, V. Patti, N. Novielli, P. Rosso, Evalita 2018: Overview on the 6th evaluation campaign of natural language processing and speech tools for italian, EVALITA Evaluation of NLP and Speech

Tools for Italian (2018) 3–8. URL: http://dx.doi.org/10.4000/books.aaccad. doi:10.4000/books.aaccademia.4437.

[12] R. Kumar, B. Lahiri, A. Ojha, Aggressive and offensive language identification in hindi, bangla, and english: A comparative study, SN Computer Science 2 (2021). doi:10.1007/s42979-020-00414-6.

[13] C. Bosco, F. Dell'Orletta, F. Poletto, M. Sanguinetti, M. Tesconi, Overview of the evalita 2018 hate speech detection task, EVALITA Evaluation of NLP and Speech Tools for Italian (2018) 67–74. URL: http://dx.doi.org/10.4000/books.aaccademia.4503. doi:10.4000/books.aaccademia.4503.

[14] A. Arango, J. Pérez, B. Poblete, Cross-lingual hate speech detection based on multilingual domain-specific word embeddings, CoRR abs/2104.14728 (2021). URL: https://arxiv.org/abs/2104.14728. arXiv:2104.14728.

[15] P. Fortuna, S. Nunes, A Survey on Automatic Detection of Hate Speech in Text, ACM Computing Surveys 51 (2019) 1–30. URL: http://dx.doi.org/10.1145/3232676. doi:10.1145/3232676.

[16] C. Corazza, S. Menini, E. Cabrio, S. T. S. Villata, Cross-Platform Evaluation for Italian Hate Speech Detection, Le Centre pour la Communication Scientifique Directe - HAL - Université de Nantes (2019).

[17] I. Markov, I. Gevers, W. Daelemans, An Ensemble Approach for Dutch Cross-Domain Hate Speech Detection, Natural Language Processing and Information Systems (2022) 3–15. URL: http://dx.doi.org/10.1007/978-3-031-08473-7_1. doi:10.1007/978-3-031-08473-7/1.

[18] D. Njagi, Z. Zuping, D. Hanyurwimfura, J. Long, A lexicon-based approach for hate speech detection, International Journal of Multimedia and Ubiquitous Engineering 10 (2015) 215–230. doi:10.14257/ijmue.2015.10.4.21.

[19] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, Proceedings of the Eleventh International Conference on Web and Social Media (2017) 512–521. URL: http://dx.doi.org/10.5555/3290605.3300749. doi:10.5555/3290605.3300749.

[20] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[21] F. Celli, M. Lai, A. Duzha, C. Bosco, V. Patti, Policycorpus XL: an italian corpus for the detection of hate speech against politics, in: E. Fersini, M. Passarotti, V. Patti (Eds.), Proceedings of the Eighth Italian Conference on Computational Linguistics, CLiC-it 2021, Milan, Italy, January 26-28, 2022, volume 3033 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021. URL: https://ceur-ws.org/Vol-3033/paper38.pdf.

[22] A. Ramponi, B. Testa, S. Tonelli, E. Jezek, Addressing religious hate online: from taxonomy creation to automated detection, PeerJ Comput. Sci. 8 (2022) e1128. URL: https://doi.org/10.7717/peerj-cs.1128. doi:10.7717/peerj-cs.1128.

[23] Twitter, Documentation, Twitter Developer Documentation, 2023. URL: https://developer.twitter.com/en/docs, Accessed: 13th June 2023.

[24] S. J. Pan, Q. Yang, A Survey on Transfer Learning, IEEE Transactions on Knowledge and Data Engineering 22 (2010) 1345–1359. URL: http://dx.doi.org/10.1109/tkde.2009.191. doi:10.1109/tkde.2009.191.

[25] L. Eric, R. Saini, G. Kovács, K. Murphy, TheNorth @ HaSpeeDe 2: BERT-based Language Model Fine-tuning for Italian Hate Speech Detection, EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020 (2020) 142–147. URL: http://dx.doi.org/10.4000/books.aaccademia.6989. doi:10.4000/books.aaccademia.6989.

[26] F. Tamburini, How "BERTology" Changed the State-of-the-Art also for Italian NLP, Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it 2020 (2020) 415–421. URL: http://dx.doi.org/10.4000/books.aaccademia.8920. doi:10.4000/books.aaccademia.8920.

[27] D. Nozza, F. Bianchi, G. Attanasio, HATE-ITA: Hate Speech Detection in Italian Social Media Text, Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH) (2022). URL: http://dx.doi.org/10.18653/v1/2022.woah-1.24. doi:10.18653/v1/2022.woah-1.24.

[28] F. He, T. Liu, D. Tao, Control batch size and learning rate to generalize well: Theoretical and empirical evidence, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 32, Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/dc6a70712a252123c40d2adba6a11d84-Paper.pdf.

[29] S. E. Ada, E. Ugur, H. L. Akin, Generalization in transfer learning, CoRR abs/1909.01331 (2019). URL: http://arxiv.org/abs/1909.01331. arXiv:1909.01331.

[30] L. Shao, F. Zhu, X. Li, Transfer learning for visual categorization: A survey, IEEE Transactions on Neural Networks and Learning Systems 26 (2015) 1019–1034. doi:10.1109/TNNLS.2014.2330900.

HODI at EVALITA 2023: Overview of the first Shared Task on Homotransphobia Detection in Italian

Debora Nozza[1], Alessandra Teresa Cignarella[2,3], Greta Damo[1], Tommaso Caselli[4] and Viviana Patti[2]

[1]Department of Computing Sciences, Bocconi University, Milan, Italy
[2]Department of Computer Science, University of Turin, Turin, Italy
[3]aequa-tech, Turin, Italy
[4]Center for Language and Cognition, University of Groningen, Groningen, The Netherlands

Abstract
HODI is a new shared task for the automatic detection of homotransphobia in Italian presented at EVALITA 2023. The challenge is organized into two subtasks: Subtask A focuses on the binary textual classification of homotransphobic tweets, while Subtask B is concerned with the identification of "rationales" for explainability in the form of textual spans of text. We have received a total of 19 runs for Subtask A and 5 runs for Subtask B from a total of 8 participating teams from 6 different countries. We present here an overview of the HODI shared task, the datasets, the evaluation methodology, the results obtained by the participants, and a discussion of the methodology adopted by the teams.

***Warning**: This paper contains examples of potentially offensive content.*[1]

Keywords
Natural Language Processing, Hate Speech, Homotransphobia

1. Introduction

Despite advancements in human and civil rights, the Internet remains a hostile environment for LGBTQIA+ individuals. The increasing frequency, severity, and complexity of online hate crimes are mirrored in the real world. In a recent ISTAT-UNAR survey[1] on discrimination on work places suffered by LGBTQIA+ people, 43.9% of the participants has been target of insults, 61.8% suffered a micro-aggression (including hate speech), and 1.1% has been physically assaulted. In addition to this, anti-LGBTQIA+ hate crimes have risen drastically in the past three years.[2] Natural Language Processing (NLP) is a key subject of research for combating online hate speech since it can automate the process at scale while reducing online moderators' labor and mental stress [2]. Despite the NLP community's interest in hate speech detection datasets and models [3], very few studies covered hate speech against the LGBTQIA+ community [4, 5, 6]. Given the *target-oriented* nature of hate speech and the ineffectiveness of transferring hate speech detection models to different unseen hate speech targets [7, 8, 9, 10], the lack of a dedicated benchmark was a pending issue for homotransphobia in Italian, which we addressed with this work.

The Homotransphobia Detection in Italian (HODI)[3] shared task at EVALITA 2023 [11] **identifies Italian hate speech directed at the LGBTQIA+ community**. This will allow us to investigate a phenomenon that has received little attention from the worldwide NLP community and has never been investigated for Italian.

Being able to automatically determine whether a message is hateful or not is an important contribution to the fight against homotransphobia, yet this is not sufficient. Systems are always prone to errors, even the most accurate ones. This means that it could be the case that a perfectly normal message gets labeled as hateful simply because it contains an identity term (e.g., the word "gay"). Or the opposite can happen. Flagging a message for this kind of content should always be accompanied by an explanation that will shed (some) light on the way the system has taken its decision. Furthermore, recent Euro-

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
debora.nozza@unibocconi.it (D. Nozza); alessandrateresa.cignarella@unito.it (A. T. Cignarella); greta.damo@studbocconi.it (G. Damo); t.caselli@rug.nl (T. Caselli); viviana.patti@unito.it (V. Patti)
https://deboranozza.com/ (D. Nozza); https://www.unito.it/persone/acignare (A. T. Cignarella); https://www.rug.nl/staff/t.caselli/ (T. Caselli); https://www.unito.it/persone/vpatti (V. Patti)
0000-0002-7998-2267 (D. Nozza); 0000-0002-4409-6679 (A. T. Cignarella); 0000-0003-2936-0256 (T. Caselli); 0000-0001-5991-370X (V. Patti)
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

[1]https://www.istat.it/it/files//2022/03/REPORTDISCRIMINAZIONILGBT_2022_rev.pdf
[2]https://www.theguardian.com/world/2021/dec/03/recorded-homophobic-hate-crimes-soared-in-pandemic-figures-show
[3]Task website: https://hodi-evalita.github.io/ and task repository: https://github.com/HODI-EVALITA/HODI_2023

CEUR-WS.org/Vol-3473/paper26.pdf

Text	Subtasks	
	A	B
Odio i fr*ci	1	[0,1,2,3,7,8,9,10,11]
Morte ai gay torinesi	1	[0,1,2,3,4,9,10,11]
Divento fr*cio per te	0	0
Gay ed etero, stessi diritti	0	0

Table 1
Examples of the annotated data.

	Subtask A		Subtask B	
Split	Hate	Not	Single Token	Multi Token
Train	2,008	2,992	48	1,960
Test	511	489	16	495

Table 2
HODI data statistic overview.

pean legislation (General Data Protection Regulation – GDPR [12]) has introduced a "right to explanation". This necessitates a paradigm change from performance-based models to interpretable models [13]. This shared task will also contribute towards this need by assessing the **models' explanation abilities** to recognize the terms relevant for hate speech. This will allow, in the future, to control for possible biases of models overfitting to specific terms (e.g., *gay*) [14, 6], as well as use the explanations to generate counternarratives.

2. Task Description

HODI is structured on two subtasks (see examples in Table 1):

- **Subtask A - Homotransphobia detection**: this is a binary classification task where systems must classify a message as hateful or not against LGBTQIA+ community.
- **Subtask B - Explainability**: once a message is classified as hateful, the objective is to identify the rationales of the classification model, i.e., those tokens in the sequence that contributed to the flagging of the message.

The two tasks are strictly interconnected, but they have been run independently.

3. Training and Testing Data

Data Collection Data have been collected from Twitter using a keyword-based approach from May 1st, 2022 until August 31st, 2022. The selection is influenced by the observation that the summer months coincide with the pride celebrations, leading to increased discussions and engagement on social media regarding the subjects relevant to our objective. Additionally, May 17th is recognized globally as the International Day Against Homophobia, Biphobia, and Transphobia, further emphasizing the significance of this time frame for our task. We focused both on keywords that are commonly used in hateful contexts (e.g., *fr*cio*) and on others related to specific events that directly involve or affect the LGBTQIA+ community (e.g., *Pride*, *DDL Zan*). The complete list of keywords can be found in Appendix A. The decision to use keywords identifying events has been done because of a tendency to observe a surge in homotransphobic messages around them. In this way, we limited the presence of only explicit profanity-driven keywords that may introduce biases in the data and, consequently, in the trained models. As a result, the final dataset does not correspond to the natural distribution of hate on social media, which is lower.

Data Annotation Our annotation guidelines[4] have been developed by re-using previous guidelines for similar shared tasks, namely HatEval [15] and AMI [16]. In particular, we define a message as being hateful by applying the following definition:

> *any communication that disparages a person or a group on the basis of some characteristics, such as color, race, ethnicity, gender, sexual orientation, religion, nationality, or other aspects.*

Following the proposals in [17], our definition of hate speech and annotation guidelines have benefited from a series of interactions with some members of the Italian LGBTQIA+ community. In addition to this, we managed to have the data manually labeled by three members of the Italian LGBTQIA+ community (two males and one female). Each message has been annotated in parallel by each annotator for both subtasks. The annotators labeled whether the text is hateful or not and targets the LGBTQIA+ community. Then, the annotation for Subtask B targeting explainability is performed following the approach in [13]. In particular, our annotators have been asked to highlight the span of text that could support their labeling decision, the so-called *rationales*. We asked annotators to provide rationales only for the tweets considered hateful. These span annotations help us to investigate deeper the manifestations of hateful speech.

[4]Available for consultation here: https://github.com/HODI-EVALITA/HODI_2023

	step1	step2	step3	*avg*
Subtask A	0.543	0.658	0.547	0.583
Subtask B	0.617	0.627	0.700	0.648

Table 3
Inter-annotator agreement calculated with Fleiss' kappa coefficient (Subtask A) and % observed agreement (Subtask B) in the three steps.

The annotation campaign has been conducted in three different steps by giving the annotators 2,000 tweets each for each step. The inter-annotator agreement (IAA) has been calculated at the end of every step. In Table 3, we display the measures of the IAA on both subtasks, calculated with Fleiss' kappa coefficient (Subtask A) and % observed agreement (Subtask B). The average of the IAA obtained in both subtasks is *substantial* according to the interpretation of [18]. It is particularly impressive how the three annotators reached an IAA of 0.648 on the selection of homotransphobic spans of text, considering the difficulty and subjectivity of the task.

Extracting Gold Labels In this shared task, we decided to provide the participants with aggregated gold labels for both tasks rather than releasing the annotations separately. The aggregation process has been implemented as follows: for Subtask A, the gold label was chosen through a majority voting strategy. Since the annotators were three, and they could select only between two labels (0/1), there was always a clear prevalence for one or the other. On the other hand, for Subtask B, the gold span of text has been established by merging the three spans selected by the three annotators. Finally, in the fashion proposed in the SemEval 2021 shared task of toxic spans detection [25], we released the annotation of spans as a list of indices referring to the position of characters in the text (see Table 1).

Data Statistics Table 2 presents a summary of the annotated data for both subtasks. We provided 5,000 training and 1,000 testing tweets. The data we provided are roughly balanced (40% hateful tweets in training and 51% in the test set). For Subtask B, we report the number of messages with a single-token rationale and those with multi-token rationales. It can be seen how in both train and test, the majority of spans containing homophobic expressions are composed of more than one token. On the other hand, in the train set, there are 48 tweets where the hateful span contains only one word. In the test set, those cases are even fewer, i.e., only 16. Table 1 shows examples of data annotations for both Subtask A and B, with the rationales highlighted in yellow for better understanding.

4. Evaluation Measures and Baseline

Systems have been evaluated using the following metrics per task:

Subtask A. We use standard evaluation metrics for text classification, namely Precision, Recall, and F1-score per class. The ranking of the systems is based on the macro-averaged F1-score of the hateful and non-hateful messages.

Subtask B. Systems are evaluated using Intersection-Over-Union (IOU) [26], an *agreement* metrics. Token-level IOU is the size of the overlap of the character of the tokens they cover divided by the size of their union. We count a prediction as a match if it overlaps with any of the ground truth rationales by more than some threshold. We use these partial matches to calculate an F1 score and subsequently rank the systems.

Two different methods have been implemented to compare models to baselines:

Subtask A. Logistic Regression classifier based on TF-ID using unigrams and bigrams only.

Subtask B. A random classifier following the implementation of the organizers of the SemEval-2021 Task 5, Toxic Spans Detection [25].

The HODI GitHub repository[5] contains the code for calculating evaluation metrics and producing predictions using the baselines.

5. Participants and Results

We have received submissions from eight teams, for a total of 18 runs for Subtask A and four for Subtask B. Only two teams participated in Subtask B. Two teams used the same approach and system architecture for participating in other EVALITA 2023 tasks, namely `O-Dang` for HaSpeeDe and `extremITA` for all tasks. The majority of the teams were from academia, with only one industrial participant.

Participants were allowed to submit a maximum number of three runs for each subtask. Note that, in the case of submissions for both tasks, participants were asked to submit their predictions for Subtask A and Subtask B at the same time, i.e., in the same evaluation window. Table 4 provides a summary of the teams, illustrating their country and the subtasks they addressed.

[5] https://github.com/HODI-EVALITA/HODI_2023

Team	Country	Task	dbmdz-BERT-Italian	AlBERTO	Open AI Davinci	IT5	Camoscio	UmBERTO - Oscar Corpus	Twitter XML-R Sentiment	Fine-tuning	Knowledge injection	Data augmentation	Multi-task Learning	Few-shot Learning	Feature Extraction	Prompting
DH-FBK [19]	IT	A, B	✓							✓		✓	✓			
CHILab [20]	IT	A		✓											✓	
extremITA [21]	IT	A, B				✓	✓									✓
O-Dang [22]	IT,UK	A		✓	✓					✓	✓					
LCTs [23]	ES,NL	A						✓		✓						
Team_Tamil [24]	IE,IN	A							✓					✓		

Table 4
Overview of the participating systems who submitted the report. We list each team's tasks, pre-trained language models, and methods investigated.

Subtask A - Homotransphobia detection The homotransphobia detection task received 19 submissions from 8 teams, as shown in Table 5. The best result has been obtained by LCTs, where the team fine-tuned an Italian pretrained RoBERTa model named UmBERTo[6] for 10 epochs. Thus, this underscores the fact that relying solely on domain-specific approaches is still insufficient when it comes to effectively utilizing large models and extensive training. 6 out of 8 teams provide better results than the baseline. Due to a code error in the official submission that was not ranked in the shared task's official results, the team CHILab resubmitted amended runs (**) after the deadline.

Subtask B - Explainability The subtask related to the identification of the rationales behind prediction decisions received 5 runs from 2 teams. Table 6 shows the results in terms of F1. Considering the task's inherently complex and unique nature, teams had to invest additional effort beyond what is typically required for a binary prediction task, leading to an anticipated decrease in participation. Both teams outperformed the random baseline. The best performing submission by extremITA obtained the homophobic rationales interrogating an instruction-tuned decoder-only model (i.e., LLaMA) with the natural language instruction "*Con quali parole l'autore del testo precedente esprime odio omotransfobico? Separa le sequenze di parole con [gap]*" (*en: In what words does the author of the previous text express homotransphobic hatred? Separate the word sequences with [gap].*). While the ability to prompt such models has already been demonstrated to be effective by [27], the Subtask B results further highlight the power of large language models to perform even more difficult subjective tasks, such as explaining homophobic hatred.

6. Discussion

In Table 4, we present an overview of the participating systems for which we have received a system description paper. This section delves into the team's varied approaches from different perspectives.

Language Models Following a trend already seen in other evaluation campaigns [15, 16], all of the proposed systems make use of pre-trained language models (PTLMs) based on encoders only (dbmdz-BERT-italian[7], AlBERTO [28], UmBERTO[8], and Twitter-XML-R-sentiment[9]), or decoders only (Open AI Davinci [29], Camoscio[10]), or using a full Transformer architecture (IT5 [30]). Only two teams used multilingual models (Twitter-XML-R-sentiment and Open AI Davinci), while all the others used Italian monolingual PTLMs. For the Italian PTLMs, only AlBERTO [28] has been trained with a language variety compatible with the task's data, i.e., social media data. It is remarkable that pure fine-tuning of PTLMs has been done only by one team (LCTs). Another team, Team_Tamil, proposes

[6] https://huggingface.co/Musixmatch/umberto-commoncrawl-cased-v1

[7] https://huggingface.co/dbmdz/bert-base-italian-cased

[8] https://huggingface.co/Musixmatch/umberto-commoncrawl-cased-v1

[9] https://huggingface.co/citizenlab/twitter-xlm-roberta-base-sentiment-finetunned

[10] https://github.com/teelinsan/camoscio

Team	Macro F1	Rank
LCTs$_3$	**0.8108**	1
LCTs$_2$	0.8000	2
O-Dang$_1$	0.7959	3
DH-FBK$_1$	0.7950	4
extremITA$_2$	0.7942	5
O-Dang$_2$	0.7920	6
DH-FBK$_2$	0.7837	7
O-Dang$_3$	0.7804	8
LCTs$_1$	0.7709	9
CHILab$_2$**	0.7525	-
CHILab$_3$**	0.7454	-
extremITA$_1$	0.7431	10
CHILab$_1$**	0.7248	-
INGEOTEC$_1$	0.7153	11
Team_Tamil$_1$	0.6735	12
Baseline	0.6691	13
SOVRAG$_3$	0.6634	14
SOVRAG$_2$	0.6334	15
SOVRAG$_1$	0.6108	16
CHILab$_3$	0.5528	17
CHILab$_1$	0.5205	18
CHILab$_2$	0.5199	19

Table 5
Results for Subtask A - Homotransphobia detection. Numbers in subscripts for the team names correspond to the submitted run.

zero and few-shot learning of fine-tuned classification language models aiming at solving hate speech detection (e.g., [31]) or emotion-related tasks (e.g., [32]) in Italian and multilingual settings. For all other participants, fine-tuning represents just one component of other architectures and solutions.

Features and Additional Data No system has used external features from specialized lexical resources. Only one participant, DH-FBK, has extended the available training materials for both subtasks using synthetic data obtained with IT5. The authors have retained only the top 2,000 examples for each class as a strategy to double the size of the HODI training set per class as well as to mitigate class imbalance.

Prompting Following recent advancements in generative language models, two teams, O-Dang and extremITA, made use of prompting engineering techniques. In the case of O-Dang, prompts have been used to query the Open AI Davinci model to extract additional data concerning the names of entities of type "PERSON" that are present in the training set. The information thus obtained is concatenated to the original message as a form of knowledge injection. The extremITA team took a more radical path by addressing all EVALITA 2023

Team	F1 score	Rank
extremITA$_2$	**0.7228**	1
DH-FBK$_1$	0.7051	2
DH-FBK$_2$	0.7008	3
extremITA$_1$	0.6598	4
Baseline	0.2050	5

Table 6
Results for Subtask B - Explainability. Numbers in subscripts for the team names correspond to the submitted run.

tasks by means of prompting. They apply two different prompting approaches, compliant with the models they use (IT5 and Camoscio). The authors exploited zero-shot prompting, which means they did not give the models any examples from the training data. They only specialized the natural language instruction for the different tasks.

Interaction between Subtask A and Subtask B The only team that exploited as much as possible the interaction between the two subtasks in the design of their system is DH-FBK. The authors developed a multi-task learning architecture using the MaChAmp v2.0 toolkit [33].

7. Conclusion and Future Work

This paper introduces HODI, the first shared task on homotransphobia detection in Italian. The task aims to not only identify homotransphobic messages but also investigate the underlying reasons behind them. We have analyzed the submissions from participating teams and concluded that satisfactory results have been achieved in detecting homotransphobia in Italian. Furthermore, notable progress has been made in the explainability task, although further work is required in this area. To continue advancing in this field, future efforts should focus on constructing larger and more diverse datasets. Additionally, there is a need to enhance the detection models and improve their ability to explain the specific words or features that contribute to a hateful classification.

Acknowledgments

The work of A.T. Cignarella and V. Patti was partially funded by the International project *STERHEOTYPES - Studying European Racial Hoaxes and sterEOTYPES*, funded by the Compagnia di San Paolo and VolksWagen Stiftung under the 'Challenges for Europe' Call for Projects (CUP: B99C20000640007). The work of D. Nozza was partially funded by Fondazione Cariplo (grant No. 2020-4288, MONICA). Debora Nozza is a member of the MilaNLP group, and the Data and Marketing Insights

Unit of the Bocconi Institute for Data Science and Analysis.

A special mention also to the people who helped us with the annotation of the dataset and the assessment of guidelines: Davide, Greta, and Mauro, thank you very much for your great help.

References

[1] D. Nozza, D. Hovy, The state of profanity obfuscation in natural language processing scientific publications, in: Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 3897–3909. URL: https://aclanthology.org/2023.findings-acl.240.

[2] M. Chaudhary, C. Saxena, H. Meng, Countering online hate speech: An NLP perspective, arXiv preprint arXiv:2109.02941 (2021). URL: https://arxiv.org/abs/2109.02941.

[3] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti, Resources and benchmark corpora for hate speech detection: a systematic review, Language Resources & Evaluation 55 (2021) 477–523.

[4] B. R. Chakravarthi, R. Priyadharshini, T. Durairaj, J. P. McCrae, P. Buitaleer, P. K. Kumaresan, R. Ponnusamy, Findings of the shared task on Homophobia Transphobia Detection in Social Media Comments, in: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, Association for Computational Linguistics, 2022.

[5] B. R. Chakravarthi, R. Priyadharshini, R. Ponnusamy, P. K. Kumaresan, K. Sampath, D. Thenmozhi, S. Thangasamy, R. Nallathambi, J. P. McCrae, Dataset for identification of homophobia and transophobia in multilingual youtube comments, arXiv preprint arXiv:2109.00227 (2021). URL: https://arxiv.org/abs/2109.00227.

[6] D. Nozza, F. Bianchi, A. Lauscher, D. Hovy, Measuring harmful sentence completion in language models for LGBTQIA+ individuals, in: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 26–34. URL: https://aclanthology.org/2022.ltedi-1.4. doi:10.18653/v1/2022.ltedi-1.4.

[7] D. Nozza, Exposing the limits of zero-shot cross-lingual hate speech detection, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics, Online, 2021, pp. 907–914. URL: https://aclanthology.org/2021.acl-short.114. doi:10.18653/v1/2021.acl-short.114.

[8] E. W. Pamungkas, V. Patti, Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Association for Computational Linguistics, Florence, Italy, 2019, pp. 363–370. URL: https://aclanthology.org/P19-2051. doi:10.18653/v1/P19-2051.

[9] E. W. Pamungkas, V. Basile, V. Patti, Misogyny detection in twitter: a multilingual and cross-domain study, Information Processing & Management 57 (2020) 102360. URL: https://www.sciencedirect.com/science/article/pii/S0306457320308554. doi:https://doi.org/10.1016/j.ipm.2020.102360.

[10] P. Chiril, E. W. Pamungkas, F. Benamara, V. Moriceau, V. Patti, Emotionally informed hate speech detection: A multi-target perspective, Cognitive Computation 14 (2022) 322–352. URL: https://doi.org/10.1007/s12559-021-09862-5. doi:10.1007/s12559-021-09862-5.

[11] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[12] Eu regulation 2016/679 general data protection regulation (GDPR), https://eur-lex.europa.eu/eli/reg/2016/679/oj, 2016. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj, accessed:2022-10-09.

[13] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, A. Mukherjee, Hatexplain: A benchmark dataset for explainable hate speech detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35:17, 2021, pp. 14867–14875.

[14] L. Dixon, J. Li, J. Sorensen, N. Thain, L. Vasserman, Measuring and mitigating unintended bias in text classification, in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 67–73. URL: https://doi.org/10.1145/3278721.3278729. doi:10.1145/3278721.3278729.

[15] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, M. Sanguinetti, SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp.

54–63. URL: https://aclanthology.org/S19-2007. doi:10.18653/v1/S19-2007.

[16] E. Fersini, D. Nozza, P. Rosso, AMI @ EVALITA2020: Automatic Misogyny Identification, in: V. Basile, D. Croce, M. Di Maro, L. C. Passaro (Eds.), Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020), CEUR.org, Online, 2020.

[17] T. Caselli, R. Cibin, C. Conforti, E. Encinas, M. Teli, Guiding principles for participatory design-inspired natural language processing, in: Proceedings of the 1st Workshop on NLP for Positive Impact, Association for Computational Linguistics, Online, 2021, pp. 27–35. URL: https://aclanthology.org/2021.nlp4posimpact-1.4. doi:10.18653/v1/2021.nlp4posimpact-1.4.

[18] J. R. Landis, G. G. Koch, An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers, Biometrics (1977) 363–374.

[19] E. Leonardelli, C. Casula, DH-FBK at HODI: Multi-Task Learning with Classifier Ensemble Agreement, Oversampling and Synthetic Data, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[20] I. Siragusa, R. Pirrone, CHILab at HODI: A minimalist approach, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[21] C. D. Hromei, D. Croce, V. Basile, R. Basili, ExtremITA at EVALITA: Multi-Task Sustainable Scaling to Large Language Models at its Extreme, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[22] C. Di Bonaventura, A. Muti, M. A. Stranisci, O-Dang at HODI and HaSpeeDe3: A Knowledge-Enhanced Approach to Homotransphobia and Hate Speech Detection in Italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[23] D. Locatelli, L. Locatelli, LCTs at HODI: Homotransphobic Speech Detection on Italian Tweets, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[24] R. Ponnusamy, P. K. Kumaresan, K. K. Ponnusamy, C. Rajkumar, R. Priyadharshini, B. Raja Chakravarthi, Team_Tamil at HODI: Few-Shot Learning for Detecting Homotransphobia in Italian Language, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[25] J. Pavlopoulos, J. Sorensen, L. Laugier, I. Androutsopoulos, SemEval-2021 task 5: Toxic spans detection, in: Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021), ACL, 2021, pp. 59–69.

[26] J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, B. C. Wallace, ERASER: A benchmark to evaluate rationalized NLP models, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4443–4458. URL: https://aclanthology.org/2020.acl-main.408. doi:10.18653/v1/2020.acl-main.408.

[27] F. M. Plaza-del arco, D. Nozza, D. Hovy, Respectful or toxic? using zero-shot learning with language models to detect hate speech, in: The 7th Workshop on Online Abuse and Harms (WOAH), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 60–68. URL: https://aclanthology.org/2023.woah-1.6.

[28] M. Polignano, P. Basile, M. de Gemmis, G. Semeraro, V. Basile, ALBERTO: Italian BERT language understanding model for NLP challenging tasks based on tweets, in: Proceedings of the 6th Italian Conference on Computational Linguistics, CLiC-it 2019, volume 2481, CEUR Workshop Proceedings (CEUR-WS. org), CEUR-WS.org, 2019. URL: http://ceur-ws.org/Vol-2481/paper57.pdf.

[29] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, Advances in Neural Information Processing Systems 35 (2022) 27730–27744.

[30] G. Sarti, M. Nissim, IT5: Large-scale text-to-text pretraining for italian language understanding and generation, ArXiv preprint 2203.03759 (2022). URL: https://arxiv.org/abs/2203.03759.

[31] D. Nozza, F. Bianchi, G. Attanasio, HATE-ITA: Hate speech detection in Italian social media text, in: Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), Association for Computational Linguistics, Seattle, Washington (Hybrid), 2022, pp. 252–260. URL: https://aclanthology.org/2022.woah-1.24. doi:10.18653/v1/2022.woah-1.24.

[32] F. Bianchi, D. Nozza, D. Hovy, FEEL-IT: Emotion and sentiment classification for the Italian language,

in: Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Association for Computational Linguistics, Online, 2021, pp. 76–83. URL: https://aclanthology.org/2021.wassa-1.8.

[33] R. van der Goot, A. Üstün, A. Ramponi, I. Sharaf, B. Plank, Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Online, 2021, pp. 176–197. URL: https://aclanthology.org/2021.eacl-demos.22. doi:10.18653/v1/2021.eacl-demos.22.

A. Appendix

Keywords used for tweet collection.

gay, pride, lesbica, f*nocchio, fr*cio, fr*cia, b*cchinaro, c*lattone, rottinc*lo, piglianc*lo, succhiac*zzo, ciucciac*zzo, c*landa, leccaf*ga, b*liccio, b*sone, f*mminiello, p*mpinaro, effeminato, c*chino, b*caiolo, ch*cca, r*cchione, invertito, travestito, passivo, deviato, ddl zan

CHILab at HODI: A minimalist approach

Irene Siragusa[1], Roberto Pirrone[1]

[1]Dipartimento di Ingegneria @ Università degli Studi di Palermo, Viale delle Scienze, Edificio 6 90129 - Palermo

Abstract

This technical report illustrates the system developed by the CHILab team for the competition HODI at EVALITA 2023. The key idea of the method we proposed for the HODI Subtask A - Homotransphobia detection, was to develop different systems arranged as suitable combinations of Pre-Trained Language Model (PTLM) for embedding extraction, neural architectures for further elaborations over the embeddings and a classifier. In particular dense layers, LSTM, BiLSTM and Transformers were used as neural architectures. The best performing system across the ones investigated in this report was made by embeddings extracted via AlBERTo coupled with a Transformer that reaches a macro-F1 score of 0.753.

Keywords

homotransphobia detection, transformer, language model

1. Warning

This paper contains examples of potentially offensive content.[1]

2. Introduction

The increasing interest for gender-inclusive and non-discriminatory language passes through its counterpart in hate speech that it is largely spreading in social networks, particularly against the LGBTQIA+ community. The NLP community is currently involved in developing systems for hate speech detection as in MAMI (Multimedia Automatic Misogyny Identification) [2] and EDOS (Explainable Detection of Online Sexism) [3] where the focus was on detection of misogyny and sexism, but these datasets are focused neither on Italian nor in detecting hate speech against people from the LGBTQIA+ community.

This paper introduces the architecture proposed by the CHILab team for the EVALITA 2023 campaign [4], and in particular as regards the Homotransphobia Detection in Italian task (HODI Subtask A - Homotransphobia detection) [5]. The general approach relies on encoding the text into suitable word embeddings that are processed via neural architectures like LSTM, BiLSTM or Transformers. Finally, the output classifier detects the presence of homotransphobic content.

We conceived our pipelines as "minimalist" architectures. No generative models [6, 7] where considered in this respect to derive embeddings. Moreover, we decided not to use fine-tuning in our PTLMs to stress the use of light networks to be trained with low computing resources. Finally, we set up a unique approach for all the tasks we have participated in EVALITA 2023.

The paper is arranged as follows: Section 2 reports a description of our systems along with data pre-processing, while results are reported and discussed in Section 3. Concluding remarks are in Section 4.

3. Description of the system

The data set by the HODI organizers contains 6000 Italian tweets annotated accordingly to the presence of homotransphobic content. Since the training set released for the competition was made up of 5000 samples, this was randomly split in a training and validation set, using a 80-20 ratio, resulting in 4000 and 1000 samples respectively.

3.1. Pre-processing

The `[URL]` tag, mention references, and retweet notes were removed since they were not considered meaningful: in particular, mentions are referred to anonymized accounts thus they add no special information. This was done after an analysis on the most cited words and hashtags[2]. As reported in Table 1, the `[URL]` tag is the most frequent one between classes and adds no information just like the anonymized mentions that are not reported. During this analysis it was interesting to notice that the most cited words are slurs directed to LGBTQIA+ members. Although a first idea for approaching the task was to look for slurs, the direct inspection of the data set shows clearly that slurs are not a good indicator of homotransphobic content. Slurs, in fact, are widely used

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT

irene.siragusa02@unipa.it (I. Siragusa); roberto.pirrone@unipa.it (R. Pirrone)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

[1]Profanities have been obfuscated with PrOf (https://github.com/dnozza/profanity-obfuscation) [1]

[2]for this analysis all the words were reported in their lower case form

CEUR-WS.org/Vol-3473/paper27.pdf

Table 1
Word distribution statistics over the dataset divided per label.

All tweets	freq	NH tweets	freq	H tweets	freq
culo	1283	culo	937	r*cchione	596
url	995	rotto	657	ch*cca	488
r*cchione	971	url	637	url	358
rotto	912	gay	544	culo	346
gay	700	c*zzo	398	rotto	255
ch*cca	688	r*cchione	375	gay	156
c*zzo	529	fare	328	isterica	151
fare	466	caghino	287	fare	138
solo	348	solo	231	c*zzo	131
me	310	me	211	quel	131

Table 2
Hashtags distribution statistics over the dataset divided per label.

All tweets	freq	NH tweets	freq	H tweets	freq
gay	15	pride	14	gioelemagaldi	9
pride	14	prelemi	13	conte	6
prelemi	14	eurovision	11	casalino	5
eurovision	13	gay	10	gay	5
tellonym	9	tellonym	8	tortura	5
gioelemagaldi	9	pridemonth	8	attacchi	4
draghi	8	dazn	7	draghi	4
pridemonth	8	meloni	6	biohacking	3
dazn	7	omofobia	5	fronte	3
conte	7	escita	5	intelligence	3

from the LGBTQIA+ people as a self-definition method suggesting a (re-)appropriation of the term itself [8], and obviously tweets of this kind cannot be considered as homotransphobic so the slur word loses its negative connotation, as in the tweet:

> firmato una fr*cia in sessione:([3]

here the term *fr*cia* does not have any negative connotation. Therefore any word-dependent consideration about the polarization of homotransphobic speeches cannot be made, as the presence of slur words does not convey negative content, i.e. slurs cannot be regarded as representative elements for separating classes. The same considerations hold for the hashtags as reported in Table 2 where the most frequent ones are neutral words. Hence, the hashtag symbol was removed and the subsequent word was kept along with its meaning inside the tweet.

Similar considerations were made for emojis: also in this case a strong polarization in the use of emojis is not found, in particular in the ones that are more associated with disgust and hate (Table 3). Since emojis are deeply used in social media communication, they were kept. Based on the statistics reported in Table 3, the emoticons

Table 3
Emoji distribution statistics over the dataset divided per label.

All	freq	NH	freq	H	freq
😂	95	😂	45	😂	50
🤣	83	🤣	44	🤣	39
😭	23	😭	19	🤮	13
❤️	19	❤️	15	😅	9
👍	19	😍	13	👍	8
😅	17	👍	11	🌈	8
💩	16	🤔	9	❗	8
🤮	15	💩	9	😎	7
😍	14	😉	8	💩	7
😉	13	🥰	8	🤡	6

contained in the data set where manually substituted with the corresponding most frequent emoji. As an example, the ":(" emoticon was translated in "😭" even if this is not the exact correspondence. This approach does not inject bias in the data set as the different emoticons were very few, while their rough meaning is preserved thus avoiding to consider them as mere sequences of punctuation marks. No further elaboration were made over the tweets: words were not reported to their lower case form, thus allowing a more accurate extraction of

[3] signed by a queer during the exam session

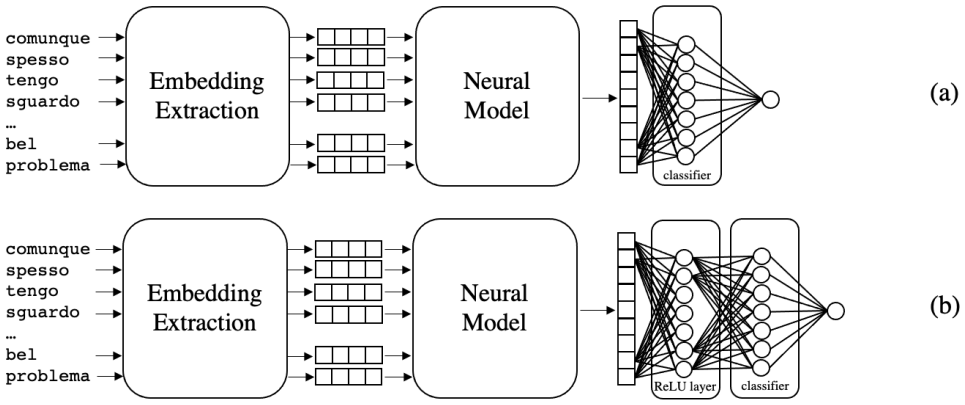


Figure 1: The first proposed architecture (a) has a module for embeddings extraction, a neural module for further processing on the extracted embeddings and a classifier. The second one (b) adds an additional ReLU dense layer.

embeddings for the case-sensitive PTLMs. As for emojis, uppercase texts has a specific meaning in social media communication in terms of prosodic and emotions interpretation [9, 10].

3.2. Network architectures

Different models were developed that share the same macro structure shown in Figure 1. The key idea was to stress, as much as possible, existent neural architectures for sequence processing, that are LSTM [11], BiLSTM and Transformers [12]. Those architecture are used to further process the extracted embeddings.

After pre-processing, the input sentences were padded to *maxLength* + 2 tokens where *maxLength* is the size of the longest sentence, and the remaining two tokens are respectively the `[CLS]` and the `[SEP]` one. Either a pre-trained language model or a static context-free embedding model were used for embedding extraction. In the last case, *fastText* [13] was used that generates a 300 tokens embedding, while a 768 tokens embedding is obtained as usual by the different PTLMs. We used the following Encoder-based Language Models in the experiments: BERT base multilingual cased [14], BERT base italian uncased [15], XLM-RoBERTa [16] and AlBERTo [17] provided by the HuggingFace Transformers library[4]. The embeddings were extracted from the last layer of the PTLMs without fine-tuning. Fine-tuning in these configuration is an option that is not taken into account since the main idea is to stress the use of light networks to be trained with low computing resources.

The extracted sequence of embeddings is fed into a neural module that consists of a LSTM or a BiLSTM or a Transformer[5]. The output feature vector has the same size of the word embedding with the exception of the BiLSTM that generates a double-length output. Finally, the feature vector is passed to a classifier made by either 300 or 768 linear units, depending on the length of the embedding, and a sigmoidal output to achieve binary classification (Figure 1.a). Some experimental configurations add an extra ReLU dense layer before the aforementioned classifier with exactly the same size. Those architectures are referred as LSTM-Deep, BiLSTM-Deep and Trasformer-Deep (Figure 1.b).

The illustrated architectures were trained only on the given data set using a machine equipped with two Intel Xeon E5 CPUs 96GB RAM and an NVIDIA TITAN Xp GPU 12GB RAM. Hyperparameters were selected as follows: dropout values in $\{0.1, 0.2\}$, batch size 32, Adam optimizer [18] with learning rate 0.01, and a Binary Cross Entropy loss. Models were trained for a maximum of 1000 epochs with a patience value of 50.

Different feature extractors were implemented using 1, 2 or 3 LSTM/BiLSTM/Transformer layers, but the best results were obtained by the single layer feature extraction modules. In addition the developed models are relative small, where the trainable parameters range from 1M to 10M.

4. Results

The best models during the evaluation window[6] were BERT-it/Transformer (run 1), AlBERTo/TransformerDeep (run 2) and AlBERTo/LSTM (run 3) and they

[4]https://huggingface.co/docs/transformers/index

[5]The corresponding architectures are named according the specific neural module

[6]they were the best models among the trained ones on our train/dev split

Table 4
The table collects the best obtained results with reference to the baseline value: the ones in italic place below the baseline; the underlined results are the ones generated removing stopwords from the data. XLM-RoBERTa, fastText and mBERT generate case-sentive embeddings.

	LSTM-Deep	**BiLSTM**	**Transformer**	**Transformer-Deep**
XLM RoBERTa	*0.565*	0.676	*0.650*	0.677
fastText	*0.554*	0.695	0.675	0.683
AlBERTo	0.717	0.745	0.705	**0.753**
mBERT	*0.333*	*0.630*	*0.338*	0.726
BERT-it	*0.642*	0.697	0.725	0.690

Table 5
The table collects the macro F1 results over the test set of the submitted models and their fixed versions (the starred ones). Result of the baseline model is also reported, along with the ranking and expected ranking position.

Run name	**Macro F1**	**Rank**
CHILab2*	0.753	10*
CHILab3*	0.745	11*
CHILab1*	0.725	13*
Baseline	0.669	13
CHILab3	0.553	17
CHILab1	0.521	18
CHILab2	0.520	19

placed ad the bottom of the rank and below the baseline [5]. Due to an internal error in the code of the training procedure, the submitted results are intrinsically wrong and for this reason, we repeated all the experiments using the correct architecture, after the release of the golden labels. An overview of all the developed models is reported in Table 4, while Table 5 shows the submitted runs, their fixed counterpart and the baseline value. In both tables the results refer to the F1-macro score over the test set and, although all possible configurations were run, in Table 4 we report the significant architecture, i.e. the configurations that placed above the baseline. The results show that the AlBERTo/Transformer architecture with a two dense layers classifier (Transformer-Deep) has the best performance, and it is expected to rank at the 10th place on the leaderboard.

Moreover, LSTM-Deep and BiLSTM models exhibit comparable performance: bi-directional sequence processing compensates for the reduced classifier's capacity. In general, the Transformer-Deep architectures performed better than the Transformer ones.

As it was expected, only the models based on *fastText* benefit from removing the stop words. The models using AlBERTo and BERT-it achieved almost the best results both in the training phase and in the evaluation, because the network can take advantage of PTLMs that are specifically fine-tuned on the target languages. In particular, AlBERTo was trained on a corpus of Italian tweets that share the same linguistic macro-structure of the data set proposed for HODI competition.

4.1. Error analysis

As suggested by the organizers of the shared tasks, an error analysis was performed particularly on the tweets that were mis-classified by the models reported in Table 4 that performs better with reference to the baseline. All classifiers agreed incorrectly on 40 tweets: the 80% of them were homotransphobic ones. Thanks to a direct analysis of their content, the following consideration can be made.

The very first consideration is that the majority of the mis-classified tweets contain slurs. As it has been shown in Section 3.1, slur words are widely used by the LGBTQIA+ people as self-reference without any discriminatory intent, so an automatic classifier may not recognize these shades of meaning as in:

> Fanculo Dolce & Gabbana non metto la roba fr*cia[7]

Moreover, many non-homotransphobic tweets share actually some linguistic similarities with the homotransphoic ones:

> DI ANORMALE c'è solo che una cripto ch*cca repressa e omofoba quale lei è #Pillon sia miserabile Senatore della Repubblica pagato dagli italiani e che peggio getta discredito sulla nostra Nazione con esternazioni quotidiane di puro, spregevole letame.[8]

Here some hateful content is reported towards a person that is considered homotransphobic. In those cases, the presence of hate speech is correctly detected but it does not meet the homotransphobic requirement.

[7] Fuck Dolce & Gabbana I do not wear f*g stuff

[8] ABNORMAL there is only that a repressed and homophobic crypto queer like you #Pillon is a miserable Senator of the Republic paid by the Italians and worse, discredits our nation with daily utterances of pure, despicable manure.

5. Conclusion

This paper reported the architectures developed by the CHILab team for HODI Subtask A promoted at the EVALITA 2023 campaign. Our models show that a relatively small classical pipeline made by embedding extraction plus further neural elaboration can have satisfactory performance in homotransphobic speech detection without the need of fine-tuning PTLMs, and using few computational resources. The use of such "minimalist" architecture is intended to allow for future development of compact explainable models where explicit linguistic knowledge is injected in the network to improve its performance.

Acknowledgments

This work is supported by the PO FESR 2014-2020 grant n. 086201000543, "SCuSi - Smart Culture in Sicily"

References

[1] D. Nozza, D. Hovy, The state of profanity obfuscation in natural language processing scientific publications, in: Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, 2023.

[2] E. Fersini, F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees, J. Sorensen, SemEval-2022 task 5: Multimedia automatic misogyny identification, in: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, Seattle, United States, 2022, pp. 533–549. URL: https://aclanthology.org/2022.semeval-1.74. doi:10.18653/v1/2022.semeval-1.74.

[3] H. R. Kirk, W. Yin, B. Vidgen, P. Röttger, SemEval-2023 Task 10: Explainable Detection of Online Sexism, in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, 2023. URL: http://arxiv.org/abs/2303.04222. doi:10.48550/arXiv.2303.04222.

[4] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[5] D. Nozza, A. T. Cignarella, G. Damo, T. Caselli, V. Patti, HODI at EVALITA 2023: Overview of the Homotransphobia Detection in Italian Task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[6] G. Mialon, R. Dessì, M. Lomeli, C. Nalmpantis, R. Pasunuru, R. Raileanu, B. Rozière, T. Schick, J. Dwivedi-Yu, A. Celikyilmaz, et al., Augmented language models: a survey, arXiv preprint arXiv:2302.07842 (2023).

[7] Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, Z. Liu, Z. Wu, D. Zhu, X. Li, N. Qiang, D. Shen, T. Liu, B. Ge, Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models, 2023. URL: http://arxiv.org/abs/2304.01852. doi:10.48550/arXiv.2304.01852, arXiv:2304.01852 [cs].

[8] E. Nossem, Queer, frocia, femminiellə, ricchione et al.–localizing "queer" in the italian context, Issue: gender/sexuality/italy, 6 (2019) (2019).

[9] M. Heath, Orthography in social media: Pragmatic and prosodic interpretations of caps lock, Proceedings of the Linguistic Society of America 3 (2018) 55–1–13. URL: https://journals.linguisticsociety.org/proceedings/index.php/PLSA/article/view/4350. doi:10.3765/plsa.v3i1.4350.

[10] S. Chan, A. Fyshe, Social and emotional correlates of capitalization on Twitter, in: Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, Association for Computational Linguistics, New Orleans, Louisiana, USA, 2018, pp. 10–15. URL: https://aclanthology.org/W18-1102. doi:10.18653/v1/W18-1102.

[11] S. Hochreiter, J. Schmidhuber, Long Short-Term Memory, Neural Computation 9 (1997) 1735–1780. URL: https://doi.org/10.1162/neco.1997.9.8.1735. doi:10.1162/neco.1997.9.8.1735. arXiv:https://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017. arXiv:1706.03762.

[13] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, T. Mikolov, Learning word vectors for 157 languages, in: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018), 2018.

[14] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[15] S. Schweter, Italian bert and electra models, 2020. URL: https://doi.org/10.5281/zenodo.4263142.

doi:10.5281/zenodo.4263142.

[16] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, CoRR abs/1911.02116 (2019). URL: http://arxiv.org/abs/1911.02116. arXiv:1911.02116.

[17] M. Polignano, P. Basile, M. de Gemmis, G. Semeraro, V. Basile, AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets, in: Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019), volume 2481, CEUR, 2019. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85074851349&partnerID=40&md5=7abed946e06f76b3825ae5e294ffac14.

[18] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

Team_Tamil at HODI: Few-Shot Learning for Detecting Homotransphobia in Italian Language

Rahul Ponnusamy[1], Prasanna Kumar Kumaresan[1], Kishore Kumar Ponnusamy[2], Charmathi Rajkumar[3], Ruba Priyadharshini[4] and Bharathi Raja Chakravarthi[1]

[1]Insight SFI Research Centre for Data Analytics, University of Galway, Ireland
[2]Guru Nanak College, Tamil Nadu, India
[3]The American College, Tamil Nadu, India
[4]Gandhigram Rural Institute-Deemed to be University, Tamil Nadu, India

Abstract

This paper presents the novel solution to HODI (Homotranphobia Detection in Italian; [1] at EVALITA 2023 [2]. The task is structured into two subtasks: Task A, homophobic message detection, and Task B, identification of rationales of homophobic messages. We participated in Task A, a binary class classification problem. The main aim of this task is to identify the homotransphobia in Italian tweets. To determine the homotransphobia, we choose the selective models that are available in the huggingface[1] related to our task. We use the zero-shot technique to select the models that are working well for the homotransphobia classification task. With those models, we performed a series of few-shot learning experiments. Our best approach achieves a macro F1-score of 0.673, higher than the baseline, and ranking number 6.

Keywords

Homotranphobia identification, Zero-shot, Few-shot learning, Text classification,

1. Introduction

Hate speech refers to any type of offensive material, including verbal, nonverbal, symbolic, or communicative actions that are intentionally used to demean members of a specific social group based on their membership [3]. Hate speech on social media is a pervasive phenomenon affecting diverse categories of users targeted because of their race, ethnicity, gender, religion, sexual orientation, political views, or other characteristics [4, 5, 6, 7]. Significant effort has been expended to combat harmful and abusive content in the general domain, primarily via reactive measures [8, 9, 10, 11, 12, 13].

One such form of hate is hate speech towards vulnerable LGBTQ+ individuals. Homophobia, a term frequently used to characterize hostile responses to lesbians and gay men, connotes a one-dimensional conception of attitudes as manifestations of irrational fears [14]. Parents, instructors, school administrators, and government are gravely concerned about homophobic bullying against gay, lesbian, bisexual, transgender, and "queer" (LGBTQ+) minorities [15]. Bullying is a form of aggression that involves a victim, an aggressor, and bully victims. Homophobia, the underlying attitude influencing bullying against LGBTQ+ vulnerable minorities, is defined as the negative beliefs, attitudes, stereotypes, and behaviors directed toward sexual minorities [16, 17].

Cross-lingual zero and few-shot learning, in addition to other detection approaches that work well with limited or nonexistent training data sets in the target language, have not been extensively studied in the current literature on homophobia [18, 19]. Neither have other detection methods that work successfully with limited or nonexistent training data sets in the target language [20]. The researchers [21] describe zero-shot learning as an extreme form of transfer learning. When this concept is used in natural language processing (NLP), a model that has been trained on one language or domain can learn to predict samples from an unseen language or domain by making use of the latent structures of a pre-trained language model that is aligned across several languages [22, 23]. In the method of cross-lingual few-shot learning, training on the source language is supplemented with samples of the target language. This helps to strengthen both cross-lingual and task-specific alignment.

In this paper, we describe a novel approach for detecting homotransphobia in Italian-language tweets shared task conducted by [1]. We have plenty of pretrained and fine-tuned (trained for a specific task) models available for the text classification tasks. We selected twelve

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
rahul.ponnusamy@insight-centre.org (R. Ponnusamy);
prasanna.kumaresan@insight-centre.org (P. K. Kumaresan);
kishorep161002@gmail.com (K. K. Ponnusamy);
charmathirajkumar@gmail.com (C. Rajkumar);
rubapriyadharshini.a@gmail.com (R. Priyadharshini);
bharathi.raja@insight-centre.org (B. R. Chakravarthi)
0000-0001-8023-7742 (R. Ponnusamy); 0000-0003-2244-246X (P. K. Kumaresan); 0000-0001-9621-668X (K. K. Ponnusamy); 0000-0003-2323-1701 (R. Priyadharshini); 0000-0002-4575-7934 (B. R. Chakravarthi)
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper28.pdf

models which are related to our task. We performed the zero-shot technique to find the model's capability on our task. It is a technique used to infer a model by only specifying the labels of our task. We selected the top three models that gave high macro F1 scores. We performed a few-shot learning technique using SetFit (Sentence Transformer Fine-tuning) [24] framework. It is a technique to train a model with few samples. Using our approach, the twitter-xlm-roberta-base-sentiment-finetunned model got the top macro F1 of 0.74 on the validation set among the model we selected from the huggingface. Our model achieves macro F1 of 0.67345 on the test set, which is higher than the baseline.

2. Related Work

The modern forms of social media are frequently exploited in ways that promote the dissemination of violent messages and remarks as well as hate speech [25]. By analyzing the people's interaction on these issues through posts, videos, and comments, a number of works have been done with the purpose of determining whether or not aggressiveness [26], misogyny [27], racism [28], harassment, and violence are present in social media [29]. On the other hand, the amount of study that has been conducted to identify homophobic and transphobic speech online has been rather limited to few research [18].

Chakravarthi et al. [30] introduced a newly developed hierarchical taxonomy for online homophobia and transphobia, as well as a dataset that has been classified by subject matter experts, that will make it possible for homophobic and transphobic content to be automatically identified. The annotators are provided with thorough annotation criteria because this is a delicate issue. The dataset includes 15,141 comments written in English, Tamil, and Tamil-English, each of which has been annotated. From this data, Chakravarthi et al. [19] organized a shared task to increase research in homophobia/transphobia identification. It garnered 10 systems for the Tamil language, 13 systems for the English language, and 11 systems for the Tamil-English language combination. The best systems for Tamil, English, and Tamil-English each received an average macro F1-score of 0.570, 0.870, and 0.610, respectively.

A similar shared task for Dravidian languages homophobia/transphobia identification was conducted by 10.1145/3574318.3574347. It is conducted in 4 language settings (Tamil, English, Malayalam, Tamil-English). It obtained 8 systems for the Tamil language, 8 systems for the English language, 9 systems for the Malayalam language, and 8 systems for Tamil-English codemixed.

Wenpeng Yin and Roth [31] examined the limitations of prior research on zero-shot text classification (0SHOT-TC), the inadequacy of getting the problem and the significance of the labels, and the chaos of datasets and assessment setups. So, they benchmark 0SHOT-TC by standardizing the datasets and evaluations. They also provided a textual entailment framework that can function with or without the annotated data of observed labels to address the more broadly defined 0SHOT-TC. This motivated us to include the zero-shot technique in our task.

3. Dataset Description

The dataset given by the EVALITA 2023 HODI [2] shared task on Homotransphobia in Italian tweets consists of labeled data categorized into two classes: homotransphobic and not-homotransphobic. The dataset is divided into three sets: training, development, and testing, and each set is carefully constructed to maintain a comparable distribution. The distribution of the dataset is visualized in Figure 1, which provides a graphical representation of the data distribution. Additionally, the class-wise distribution of the training, development, and test sets is presented in Table 3, as provided by the HODI shared task organizers. These distributions offer valuable insights into the composition and balance of the dataset, enabling researchers to analyze and develop models to address homotransphobia in Italian tweets effectively.

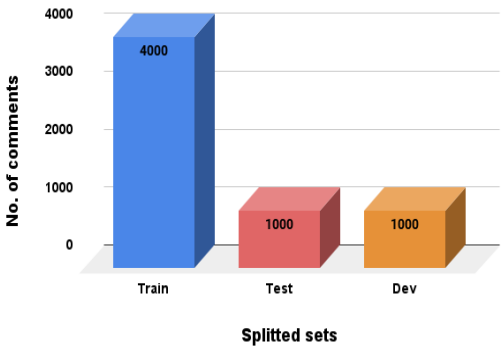


Figure 1: Visualization of splitted dataset

Further, we select 400 random samples from the training set, and we have split the dataset for the few-shot learning into five sets (80, 160, 240, 320, and 400). These sets are divided based on 20, 40, 60, 80, and 100 percent of the data taken from the training data. These datasets are visualized in Figure 2.

4. Methodology

We thoroughly describe our experimental setting in this section. All the experiments that we did adhere to the

	Labels	Comments	Total
Training	0	2379	4000
	1	1621	
Development	0	613	1000
	1	387	
Test	0	489	1000
	1	511	
	Total		**6000**

Table 1
This table shows the Data Distributions of the training, development, and test set. 1 denotes homotransphobic and 0 denotes not-homophobic

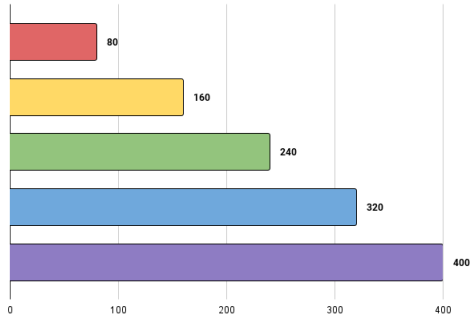


Figure 2: Split the dataset for Few-Shot learning

two phases listed below.

Zero-shot learning: In zero-shot learning, the model receives solely natural language instructions describing the task, and demonstrations are not allowed. The model is expected to generalize and perform the task based on the provided instructions without seeing explicit examples during training [32]. We took 12 models in total, which include natural language inference models and the model that trained for different tasks (sentiment, hate, emotions, toxicity) for our study. The list of models can be viewed in Table 4. These models are chosen based on topics related to the homotransphobic detection task.

Few-shot learning: It is a technique in which a limited number of task demonstrations with samples are provided to the model as training, but no weight updates are permitted [33]. With the top three models performing well in zero-shot, we applied a few-shot technique on them with the few sets of data from the training data, shown in Figure 2. We used the SetFit framework to perform few-shot text classification. Before proceeding with few-shot learning, we divided the dataset into five subsets, each containing a different number of samples: 80, 160, 240, 320, and 400. This division allowed us to examine how the models performed with increasing amounts of labeled data. We fine-tuned the models for 8 epochs, trained on GeForce GTX 1080 Ti GPU with a learning rate of 2×10^{-5} and batch size of 2. We used the same setting for all the models used for fine-tuning.

5. Results and Discussion

In this section, we show our results of the experiment and the techniques aimed to address the task given by the HODI shared task at EVALITA 2023.

Our study employed a diverse set of pre-trained models for zero-shot learning. The models used in this phase to detect the homotransphobic content in Italian encompassed a wide range of architectures and pre-trained representations, including the models that are presented in Table 4.

The evaluation of our models' performance was based on the macro F1 score, a widely used metric that considers the precision and recall across all classes. By employing the macro averaging technique, we ensured that each class contributed equally to the overall score, providing a comprehensive assessment of the models. After evaluating the models' performance, we observed notable variations in their effectiveness. Among the models, twitter-xlm-roberta-base-sentiment-finetunned, xlm-roberta-large-it-mnli[1], and mDeBERTa-v3-base-xnli-multilingual-nli-2mil7 emerged as the top performers. These models demonstrated superior performance when compared to the others, achieving macro F1 scores of 0.52, 0.51, and 0.48, respectively, in the validation set. The higher F1 scores indicate that these models effectively captured the homotransphobic information for the given task, showcasing the advantages of leveraging transfer learning techniques in zero-shot learning scenarios.

We performed the few-shot learning experiment with 5 different numbers of few samples. We observed significant performance improvements across the tested models following the few-shot learning experiments. Notably, the twitter-xlm-roberta-base-sentiment-finetunned model outperformed the others, achieving a remarkable macro F1 score of 0.74 on the validation set when trained on the 400-sample subset. The results for the few-shot learning are tabulated in Table 4. This model performed well and got a macro F1 of 0.6735 in the test set, which is higher than the baseline result of 0.6691. This finding suggests that the model effectively learned from the limited labeled examples in the few-shot setting, showcasing its ability to generalize and adapt to novel instances to detect the homotransphobic for the Italian.

It is important to note that while the top-performing models showed promising results in both zero-shot and few-shot learning scenarios, further evaluation and comparison with other state-of-the-art approaches are necessary to establish their performance in a broader context.

[1] https://huggingface.co/Jiva/xlm-roberta-large-it-mnli

Models	ACC	M_R	M_F1	W_P	W_R	W_F1	M_P
twitter-xlm-roberta-base-sentiment-finetunned [34]	0.52	0.53	**0.52**	0.55	0.52	0.53	0.52
xlm-roberta-large-it-mnli	0.52	0.52	**0.51**	0.54	0.52	0.52	0.52
mDeBERTa-v3-base-xnli-multilingual-nli-2mil7 [35]	0.58	0.51	**0.48**	0.53	0.58	0.53	0.51
dehatebert-mono-italian [36]	0.45	0.49	0.45	0.51	0.44	0.44	0.49
mdeberta-v3-base-tasksource-nli [37]	0.44	0.49	0.44	0.52	0.44	0.42	0.49
feel-it-italian-emotion [38]	0.59	0.50	0.43	0.52	0.59	0.50	0.50
mmarco-mMiniLMv2-L12-H384-v1 [39]	0.42	0.44	0.42	0.46	0.42	0.42	0.44
hate_speech_it	0.59	0.51	0.42	0.54	0.59	0.52	0.52
distilbert-base-multilingual-cased-toxicity	0.61	0.50	0.39	0.59	0.61	0.48	0.58
hate-ita-xlm-r-base [40]	0.41	0.49	0.39	0.50	0.41	0.36	0.48
feel-it-italian-sentiment [38]	0.40	0.49	0.34	0.49	0.40	0.29	0.47
setfit-italian-hate-speech [24]	0.39	0.50	0.30	0.51	0.39	0.24	0.49

Table 2
Results for Zero-Shot Learning in the validation set

Models	Shots	Acc	M_P	M_R	M_F1	W_P	W_R	W_F1
mDeBERTa-v3-base-xnli-multilingual-nli-2mil7	**80**	0.65	0.63	0.62	0.63	0.65	0.65	0.65
	160	0.66	0.64	0.64	0.64	0.66	0.66	0.66
	240	0.73	0.71	0.71	0.71	0.73	0.73	0.73
	320	0.68	0.66	0.65	0.65	0.67	0.68	0.68
	400	0.72	0.71	0.69	0.70	0.72	0.72	0.72
twitter-xlm-roberta-base-sentiment-finetunned	**80**	0.68	0.66	0.63	0.63	0.67	0.68	0.66
	160	0.69	0.67	0.67	0.67	0.69	0.69	0.69
	240	0.72	0.71	0.70	0.70	0.72	0.72	0.72
	320	0.75	0.75	0.72	0.72	0.75	0.75	0.74
	400	0.76	0.74	0.75	**0.74**	0.76	0.76	0.76
xlm-roberta-large-it-mnli	**80**	0.64	0.61	0.59	0.59	0.62	0.64	0.62
	160	0.70	0.55	0.52	0.51	0.63	0.70	0.64
	240	0.61	0.31	0.50	0.38	0.38	0.61	0.47
	320	0.61	0.31	0.50	0.38	0.35	0.61	0.47
	400	0.61	0.31	0.50	0.38	0.38	0.61	0.47

Table 3
Results for Few-Shot Learning in the validation set

Additionally, future research should explore the behavior of these models with larger datasets and investigate their robustness and generalization capabilities in different domains. These additional investigations will help to provide a more comprehensive understanding of the model's strengths and limitations.

6. Conclusion

In this study, our team, TEAM_TAMIL, presented our solutions for the Homotransphobia Detection in Italian (HODI) Shared Task at EVALITA 2023. Our approach focused on utilizing pre-trained models for zero-shot learning, aiming to detect homotransphobia in Italian text. We experimented with multiple pre-trained models and evaluated their performance to identify the most effective one. Additionally, we employed a few-shot learning technique by utilizing a dataset consisting of five sets, as depicted in the dataset section diagram. Our findings revealed that the selected model from zero-shot learning performed well in the few-shot learning scenario. As a result, our model achieved the 6th position in the ranking for subtask A with macro F1 of 0.6735 in the test samples. From this, we conclude that our novel approach works better than the baseline results and can be used to improve homotransphobia identification in the future.

References

[1] D. Nozza, A. T. Cignarella, G. Damo, T. Caselli, V. Patti, HODI at EVALITA 2023: Overview of the Homotransphobia Detection in Italian Task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[2] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th

evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[3] O. Ștefăniță, D.-M. Buf, Hate speech in social media and its effects on the LGBT community: A review of the current research, Romanian Journal of Communication and Public Relations 23 (2021) 47–55.

[4] W. Warner, J. Hirschberg, Detecting hate speech on the world wide web, in: Proceedings of the Second Workshop on Language in Social Media, Association for Computational Linguistics, Montréal, Canada, 2012, pp. 19–26. URL: https://aclanthology.org/W12-2103.

[5] A. Schmidt, M. Wiegand, A survey on hate speech detection using natural language processing, in: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 1–10. URL: https://aclanthology.org/W17-1101. doi:10.18653/v1/W17-1101.

[6] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, ACM Computing Surveys (CSUR) 51 (2018) 1–30.

[7] N. Chetty, S. Alathur, Hate speech review in the context of online social networks, Aggression and violent behavior 40 (2018) 108–118.

[8] A. Founta, L. Specia, A survey of online hate speech through the causal lens, in: Proceedings of the First Workshop on Causal Inference and NLP, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 74–82. URL: https://aclanthology.org/2021.cinlp-1.6. doi:10.18653/v1/2021.cinlp-1.6.

[9] P. K. Kumaresan, Premjith, R. Sakuntharaj, S. Thavareesan, S. Navaneethakrishnan, A. K. Madasamy, B. R. Chakravarthi, J. P. McCrae, Findings of Shared Task on Offensive Language Identification in Tamil and Malayalam, in: Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE '21, Association for Computing Machinery, New York, NY, USA, 2022, p. 16–18. URL: https://doi.org/10.1145/3503162.3503179. doi:10.1145/3503162.3503179.

[10] R. Priyadharshini, B. R. Chakravarthi, S. Cn, T. Durairaj, M. Subramanian, K. Shanmugavadivel, S. U Hegde, P. Kumaresan, Overview of abusive comment detection in Tamil-ACL 2022, in: Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 292–298. URL: https://aclanthology.org/2022.dravidianlangtech-1.44. doi:10.18653/v1/2022.dravidianlangtech-1.44.

[11] M. R. Karim, S. K. Dey, T. Islam, M. Shajalal, B. R. Chakravarthi, Multimodal hate speech detection from Bengali memes and texts, in: Speech and Language Technologies for Low-Resource Languages: First International Conference, SPELLL 2022, Kalavakkam, India, November 23–25, 2022, Proceedings, Springer, 2023, pp. 293–308.

[12] B. R. Chakravarthi, M. B. Jagadeeshan, V. Palanikumar, R. Priyadharshini, Offensive language identification in Dravidian languages using MPNet and CNN, International Journal of Information Management Data Insights 3 (2023) 100151. URL: https://www.sciencedirect.com/science/article/pii/S2667096822000945. doi:https://doi.org/10.1016/j.jjimei.2022.100151.

[13] B. R. Chakravarthi, R. Priyadharshini, S. Banerjee, M. B. Jagadeeshan, P. K. Kumaresan, R. Ponnusamy, S. Benhur, J. P. McCrae, Detecting abusive comments at a fine-grained level in a low-resource language, Natural Language Processing Journal 3 (2023) 100006. URL: https://www.sciencedirect.com/science/article/pii/S2949719123000031. doi:https://doi.org/10.1016/j.nlp.2023.100006.

[14] G. M. Herek, Beyond" homophobia": A social psychological perspective on attitudes toward lesbians and gay men, Journal of homosexuality 10 (1984) 1–21.

[15] S. S. Horn, J. G. Kosciw, S. T. Russell, Special issue introduction: New research on lesbian, gay, bisexual, and transgender youth: Studying lives in context, 2009.

[16] J. S. Hong, J. Garbarino, Risk and protective factors for homophobic bullying in schools: An application of the social–ecological framework, Educational Psychology Review 24 (2012) 271–285.

[17] B. R. Chakravarthi, Detection of homophobia and transphobia in youtube comments, International Journal of Data Science and Analytics (2023). URL: https://doi.org/10.1007/s41060-023-00400-0. doi:10.1007/s41060-023-00400-0.

[18] B. R. Chakravarthi, A. Hande, R. Ponnusamy, P. K. Kumaresan, R. Priyadharshini, How can we detect homophobia and transphobia? experiments in a multilingual code-mixed setting for social media governance, International Journal of Information Management Data Insights 2 (2022) 100119. URL: https://www.sciencedirect.com/science/article/pii/S2667096822000623. doi:https://doi.org/10.1016/j.jjimei.2022.100119.

[19] B. R. Chakravarthi, R. Priyadharshini, T. Durairaj, J. McCrae, P. Buitelaar, P. Kumaresan, R. Ponnusamy, Overview of the shared task on homophobia and transphobia detection in social media

comments, in: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 369–377. URL: https://aclanthology.org/2022.ltedi-1.57. doi:10.18653/v1/2022.ltedi-1.57.

[20] Y. Wang, Q. Yao, J. T. Kwok, L. M. Ni, Generalizing from a few examples: A survey on few-shot learning, ACM computing surveys (csur) 53 (2020) 1–34.

[21] I. Goodfellow, Y. Bengio, A. Courville, Deep learning, MIT press, 2016.

[22] G. Winata, S. Wu, M. Kulkarni, T. Solorio, D. Preotiuc-Pietro, Cross-lingual few-shot learning on unseen languages, in: Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online only, 2022, pp. 777–791. URL: https://aclanthology.org/2022.aacl-main.59.

[23] Y. Gu, X. Han, Z. Liu, M. Huang, PPT: Pre-trained prompt tuning for few-shot learning, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 8410–8423. URL: https://aclanthology.org/2022.acl-long.576. doi:10.18653/v1/2022.acl-long.576.

[24] L. Tunstall, N. Reimers, U. E. S. Jo, L. Bates, D. Korat, M. Wasserblat, O. Pereg, Efficient few-shot learning without prompts, arXiv preprint arXiv:2209.11055 (2022).

[25] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on Twitter, in: Proceedings of the NAACL Student Research Workshop, Association for Computational Linguistics, San Diego, California, 2016, pp. 88–93. URL: https://aclanthology.org/N16-2013. doi:10.18653/v1/N16-2013.

[26] S. T. Aroyehun, A. Gelbukh, Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling, in: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 90–97. URL: https://aclanthology.org/W18-4411.

[27] P. Zeinert, N. Inie, L. Derczynski, Annotating online misogyny, 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), Association for Computational Linguistics, Online, 2021, pp. 3181–3197. URL: https://aclanthology.org/2021.acl-long.247. doi:10.18653/v1/2021.acl-long.247.

[28] A. Field, S. L. Blodgett, Z. Waseem, Y. Tsvetkov, A survey of race, racism, and anti-racism in NLP, 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), Association for Computational Linguistics, Online, 2021, pp. 1905–1925. URL: https://aclanthology.org/2021.acl-long.149. doi:10.18653/v1/2021.acl-long.149.

[29] K. Glasgow, R. Schouten, Assessing violence risk in threatening communications, in: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, Association for Computational Linguistics, Baltimore, Maryland, USA, 2014, pp. 38–45. URL: https://aclanthology.org/W14-3205. doi:10.3115/v1/W14-3205.

[30] B. R. Chakravarthi, R. Priyadharshini, R. Ponnusamy, P. K. Kumaresan, K. Sampath, D. Thenmozhi, S. Thangasamy, R. Nallathambi, J. P. McCrae, Dataset for identification of homophobia and transophobia in multilingual youtube comments, arXiv preprint arXiv:2109.00227 (2021).

[31] J. H. Wenpeng Yin, D. Roth, Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach, in: EMNLP, 2019. URL: https://arxiv.org/abs/1909.00161.

[32] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[33] D. Nozza, Exposing the limits of zero-shot cross-lingual hate speech detection, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2021, pp. 907–914.

[34] F. Barbieri, L. Espinosa Anke, J. Camacho-Collados, XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 258–266. URL: https://aclanthology.org/2022.lrec-1.27.

[35] M. Laurer, W. v. Atteveldt, A. S. Casas, K. Welbers, Less Annotating, More Classifying – Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT - NLI, Preprint (2022). URL: https://osf.io/74b8k, publisher: Open Science Framework.

[36] S. S. Aluru, B. Mathew, P. Saha, A. Mukherjee, Deep learning models for multilingual hate speech detection, arXiv preprint arXiv:2004.06465 (2020).

[37] D. Sileo, tasksource: Structured dataset preprocessing annotations for frictionless extreme multitask learning and evaluation, arXiv preprint arXiv:2301.05948 (2023). URL: https://arxiv.org/abs/2301.05948.

[38] F. Bianchi, D. Nozza, D. Hovy, "FEEL-IT: Emotion and Sentiment Classification for the Italian Language", in: Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Association for Computational Linguistics, 2021.

[39] W. Wang, H. Bao, S. Huang, L. Dong, F. Wei, MiniLMv2: Multi-head self-attention relation distillation for compressing pretrained transformers, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 2140–2151. URL: https://aclanthology.org/2021.findings-acl.188. doi:10.18653/v1/2021.findings-acl.188.

[40] D. Nozza, F. Bianchi, G. Attanasio, HATE-ITA: Hate speech detection in italian social media text, in: Proceedings of the 6th Workshop on Online Abuse and Harms, Association for Computational Linguistics, 2022.

LCTs at HODI: Homotransphobic Speech Detection on Italian Tweets

Davide Locatelli[1], Lorenzo Locatelli[2]

[1]Technical University of Catalonia, 31 Calle Jordi Girona, 08034 Barcelona, Spain
[2]University of Groningen, Broerstraat 5, 9712 CP Groningen, Netherlands

Abstract

Recent research highlighted the importance of employing language and culture-specific techniques to accurately detect homotransphobic speech. In this paper, we present our involvement in Subtask A of EVALITA 2023's HODI shared task [1], which specifically addresses the identification of homotransphobic content in Italian tweets. Our approach employs a classifier built upon pre-trained Italian word embeddings. Our approach achieves the best results in the shared task, and can serve as a valuable tool to combat this harmful phenomenon. We release our code at https://github.com/davidelct/hodi2023.

***Warning**: This paper contains examples of potentially offensive content. Profanities have been obfuscated with PrOf (https://github.com/dnozza/profanity-obfuscation) [2]*

Keywords

hate speech detection, homotransphobia, social media

1. Introduction

Social media platforms have revolutionized communication, providing a space for diverse viewpoints and opinions to be shared. While these platforms offer invaluable means of connection and expression, they have unfortunately also become breeding grounds for online harassment, particularly targeting minorities. This pervasive issue has raised significant concerns about the safety and well-being of the LGBTQIA+ community, which often faces homotransphobic harassment in digital spaces [3].

One of the challenges associated with combating online harassment is the ease with which users can freely express prejudiced views without immediate consequences. Compounding the problem, social media algorithms often contribute to the formation of echo chambers, where individuals are predominantly exposed to content that reinforces their existing beliefs [4]. Consequently, these algorithms can inadvertently perpetuate discriminatory attitudes and create an environment where hate speech thrives.

To address this pressing problem, the field of natural language processing (NLP) offers valuable resources that can effectively identify harmful online content and reduce its prevalence through automated hate speech detection systems. By leveraging NLP techniques, online moderators, who shoulder the responsibility of identifying and flagging dangerous content, can significantly alleviate the psychological strain associated with their role.

Recent research has shed light on the pervasive and complex nature of online homotransphobic hate speech [5, 6], showing a strong correlation between such hate speech and specific linguistic and cultural factors. This emphasizes the importance of targeted strategies that consider the linguistic context in which it occurs.

In this paper, we explore a promising approach to address the issue of homotransphobic hate speech on social media. Specifically, we leverage pre-trained word embeddings derived from large language models to build a classifier.

We use Subtask A of the HODI shared task [1] from the EVALITA 2023 workshop [7] to demonstrate that a classifier based on monolingual Italian word embeddings yields high results, highlighting how this approach can capture the nuances of the cultural factors at play. In Subtask A the goal is to predict whether a given tweet contains homotransphobic speech or not. We found that our approach achieves the highest results in the shared task.

The remainder of this paper is organized as follows. Section 2 describes the data used in this work, and the preprocessing techniques we employed. Our methodology, including the specifics of our experimental setup, is presented in Section 3. Section 4 showcases the results we obtained, while Section 5 contains a qualitative analysis of the errors made by the different models in our study. Section 6 concludes the paper discussing the implications of our research and proposing future directions to tackle homotransphobic hate speech on social media.

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
davide.locatelli@upc.edu (D. Locatelli)
https://davidelct.com (D. Locatelli)
0009-0006-4194-4907 (D. Locatelli)
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper30.pdf

Table 1
Number of *homotransphobic* and *non-homotransphobic* tweets in the HODI Subtask A dataset.

Split	Homotransphobic	Non-Homotransphobic
Train	2,008	2,992
Test	511	489
Total	2,519	3,481

2. Data

Here, we present an overview of the data utilized in our study. This includes both the data released as part of the HODI challenge, as well as the data on which the models we utilized were pre-trained on. We did not undertake the pre-training step ourselves; nevertheless, we believe that describing the data is essential to offer a comprehensive understanding of the information to which the model has been exposed.

2.1. HODI Dataset

The HODI task organizers provided 6,000 Italian tweets manually labeled by expert annotators. For Subtask A, the annotators categorized tweets into two classes: homotransphobic or non-homotransphobic. The dataset was split into 5,000 tweets for training and 1,000 tweets for testing. To monitor the progress of our experiments, we reserved 200 tweets from the training set for validation.

The dataset statistics, as presented in Table 2.1, reflect a well-balanced distribution between the two classes across both the training and testing splits. This equilibrium enhances the reliability of our results and ensures that our model receives sufficient exposure to diverse instances of homotransphobic and non-homotransphobic language in Italian tweets during the fine-tuning process.

2.2. OSCAR Dataset

During the pre-training phase, the data utilized was the Italian corpus from the OSCAR dataset [8]. This particular collection of data is extensive, consisting of approximately 70GB of plain text. Specifically, it contains 210 million sentences and 11 billion words. The inclusion of such a vast amount of linguistic data ensures the model's exposure to a wide range of sentence structures, vocabulary, and syntactic patterns present in the Italian language.

3. Methodology

In this section we illustrate our approach, explaining both the data pre-processing steps we undertook, as well as the details of the models we utilized for Subtask A.

Table 2
Hyperparameters of models run1, run2 and run3

Hyperparameter	Value
N. epochs	{3, 5, 10}
Training batch size	16
Valid batch size	16
Warmup steps	500
Weight decay	0.01
Learning rate	2e-5

3.1. Data pre-processing

Our pre-processing consists of removing usernames, hashtags, and unnecessary white spaces from the tweets. To tokenize the text, we utilize the tokenizer associated with the pre-trained model that we describe in the next section.

3.2. Models

The three models used in our submission all consist of classifiers built on top of UmBERTo [9]. The three models all share the same hyperparameters (see Table 3.2), but they differ in the number of fine-tuning epochs on the HODI Subtask A data. Specifically:

Model run1 was fine-tuned for 3 epochs.

Model run2 was fine-tuned for 5 epochs.

Model run3 was fine-tuned for 10 epochs.

UmBERTo is a Roberta-base language model [10] pre-trained on Italian text using SentencePiece and Whole Word Masking techniques. For our classification tasks, we specifically utilized the `UmBERTo-Commoncrawl-Cased` version. [1] Using the HuggingFace Transformers library [11], we applied a classification head on top of the model outputs, which enabled us to fine-tune the base model on the HODI data for Subtask A.

The selection of the UmBERTo-Commoncrawl-Cased version offers enhanced compatibility with a wide array of text sources in comparison to alternative versions such as `Umberto-wikipedia-uncased-v1`. The latter model is pre-trained on a smaller dataset consisting mainly of Wikipedia posts, resulting in a narrower variety of text types compared to OSCAR. Furthermore, the version we selected retains the original casing of the text, which can provide significant insights especially in social media posts, where casing often serves as a means to convey

[1] Available at https://huggingface.co/Musixmatch/umberto-commoncrawl-cased-v1.

Table 3
Results of all model runs submitted to the HODI Subtask A. Our runs are underlined.

Team name	Model name	Macro F1 score
LCTs	run3	0.8108
LCTs	run2	0.8000
odang4hodi	run1	0.7959
DH-FBK	run1	0.7950
extremITA	run2	0.7942
odang4hodi	run2	0.7920
DH-FBK	run2	0.7837
odang4hodi	run3	0.7804
LCTs	run1	0.7709
extremITA	run1	0.7431
INGEOTEC	run1	0.7153
Team Tamil	run1	0.6735
baseline	run1	0.6691
SOVRAG	run3	0.6634
SOVRAG	run2	0.6334
SOVRAG	run1	0.6108
CHILab	run3	0.5528
CHILab	run1	0.5205
CHILab	run2	0.5199

strong emotions, opinions, and emphasis, and can prove as a valuable signal to detect hate speech.

To optimize our model, we employ the AdamW optimizer [12] and utilize a linear learning rate scheduler. In Table 3.2, we provide information about our experimental configuration, outlining the specific hyperparameters we selected.

4. Results

To assess the accuracy of our model's predictions, we employ the Macro F1 score as the evaluation metric. Table 4 reports the results of our three runs, as well as all other submissions to the HODI Subtask A.

We can observe that our approach is highly competitive in the shared task. Specifically, **Model run3** and **Model run2** achieve the highest and second-highest score in the competition, with over 0.80 Macro F1 performance. However, it should be noted that all models in the top five achieve over 0.79 Macro F1, and are within 0.2 point difference. While **Model run1** does not appear in the top five runs, it still achieves over 0.77 Macro F1.

Focusing only on our runs, it is evident that the performance improves as we extend the fine-tuning process, as demonstrated by the increment in the score with additional epochs. This observation highlights the positive impact of longer fine-tuning periods on the model's predictive capabilities. By allowing the model to undergo more epochs, we enable it to refine its predictions.

Table 4
Top words from the false negative examples across all three models.

Word	English translation	Count
F*mminiello	Effeminate gay man	20
F*mminielli	Effeminate gay men	13
Rotto	Broken	7
Culo	A*s	6
C*zzo	D*ck	5
Gay	Gay	5
C*lattone	F*dgepacker	5
R*cchione	F*ggot	4
Lesbiche	Lesbians	4
Tr*vioni	Tr*nsvestite	3

5. Error analysis

We divide the error analysis in two parts. First we consider examples that have been incorrectly categorized as not homotransphobic by all models, despite the gold label indicating the presence of homotransphobic speech. In other words, we consider false negatives across all models. This is so that we can gain an understanding of where our system would fail to protect LGBTQIA+ individuals online, highlighting directions for further refinements.

Then we analyze examples on which **Model run1** and **run2** failed to identify homotransphobia, but on which **run3** succeeded. This is to gain an understanding of the impact of extended fine-tuning.

5.1. False negatives

In total, 108 examples were false negatives across models, i.e. were wrongly classified as not homotransphobic by all three models. We report the top 10 words appearing in these examples in Table 5.1. It is interesting to note that the term "f*mminiello" and its plural form are the most frequently occurring words.

This observation is noteworthy as the word is primarily used in the Neapolitan dialect rather than being widely employed throughout Italy. It suggests that all models struggle with dialectal words that are infrequently encountered in its Italian pre-training corpus. Further investigation revealed that the fine-tuning data for HODI Subtask A only included two tweets containing such word, explaining why none of the models recognized this particular case.

The remaining words in the table consist of various slurs, such as "rotto in culo" (a combination of the third and fourth words), which translates to "assf*cked." This expression stigmatizes anal sex and, as it is predominantly used in its masculine form to insult men, it implies a negative connotation towards gay male sex. However, it is important to note that this expression is also commonly

Table 5
Top words from the tweets where model run3 improved compared to the other models.

Word	English translation	Count
Seduto	Sat down	4
F*mminielli	Effeminate gay men	3
Grandissimo	Very big	2
Figlio	Son	2
Casa	Home	2
F*mminiello	Effeminate gay man	2
GIOELE	First name (male)	2
MAGALDI	Last name	2
Problema	Problem	2
Verona	Verona (city)	2

used to insult non-gay individuals, making the identification of harassment towards LGBTQIA+ individuals more complex and context-dependent (e.g., considering the identity of the person being targeted). Nevertheless, it is worth mentioning that even when used to target non-LGBTQIA+ individuals, many people may still consider such expression to be homotransphobic, which is up for debate.

5.2. Improvements from extended fine-tuning

In total, 29 examples were correctly classified by **Model run3**, and incorrectly classified by the other two. We report the top 10 words appearing in these examples in Table 5.2.

We can observe that model run3 corrects a few of the false negatives containing the word "f*mminiello" described above, suggesting that more epochs allow the model to pick up on more subtle patterns present in the rest of the tweets.

Another interesting phenomenon is that of the words "GIOELE MAGALDI", which are a first and last name of an Italian male author, often insulted on social media with homotransphobic slurs. It is interesting to observe that model 3 was able to pick up on the harassment of an individual compared to the previous runs. This author is often insulted in all-caps tweets, which might have helped the model pick up on the aggressiveness of the language.

6. Conclusion

In this paper we described our approach to the HODI Subtask A [1] at EVALITA 2023 [7] on homotransphobic speech detection. The goal of our participation was to assess the effectiveness of using a simple classifier based on monolingual pre-trained word embeddings. We built our model on top of UmBERTo, an Italian version of BERT, pre-trained on a large amount of Italian data. We fine-tuned it using the HODI Subtask A data. We experimented by running the fine-tuning process for different number of epochs, and obtained high Macro F1 scores for all runs, around 0.8.

In future work, it would be worth comparing this performance with that of classifiers based on multilingual pre-trained word embeddings. Given the linguistic and culture-specific phenomena that characterize homotransphobic speech, it would be interesting to understand whether targeted monolingual embeddings yield better results than multilingual ones, potentially uncovering whether the former have a better time with nuanced edge cases.

While Italian is not a low-resource language, it would be also interesting to run this experiment with multilingual embeddings obtained from a dataset that does not include Italian, to understand whether the model can generalize from languages that exhibit similar phenomena as the target.

Acknowledgments

We thank the task organizers for setting up this shared challenge and providing the HODI dataset, a valuable resource for future work on this important area of research. Davide Locatelli is part of the INTERACT group of the Technical University of Catalonia, and is supported by the European Research Council under the European Union's Horizon 2020 research and innovation program (grant No. 853459). We gratefully acknowledge the computer resources at Artemisa, funded by the European Union ERDF and Comunitat Valenciana, and the technical support provided by the Instituto de Fisica Corpuscular, IFIC (CSIC-UV).

References

[1] D. Nozza, A. T. Cignarella, G. Damo, T. Caselli, V. Patti, HODI at EVALITA 2023: Overview of the Homotransphobia Detection in Italian Task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[2] D. Nozza, D. Hovy, The state of profanity obfuscation in natural language processing scientific publications, in: Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, 2023.

[3] GLAAD, Social media safety index, 2022. URL: https://sites.google.com/glaad.org/smsi/platform-scores, accessed: 2023-07-22.

[4] M. Cinelli, G. D. F. Morales, A. Galeazzi, W. Quattrociocchi, M. Starnini, The echo chamber effect on social media, Proceedings of the National Academy of Sciences 118 (2021) e2023301118. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2023301118. doi:10.1073/pnas.2023301118.

[5] B. R. Chakravarthi, R. Priyadharshini, R. Ponnusamy, P. K. Kumaresan, K. Sampath, D. Thenmozhi, S. Thangasamy, R. Nallathambi, J. P. McCrae, Dataset for identification of homophobia and transophobia in multilingual youtube comments, ArXiv abs/2109.00227 (2021).

[6] D. Locatelli, G. Damo, D. Nozza, A cross-lingual study of homotransphobia on Twitter, in: Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP), Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 16–24. URL: https://aclanthology.org/2023.c3nlp-1.3.

[7] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[8] P. J. Ortiz Suárez, L. Romary, B. Sagot, A monolingual approach to contextualized word embeddings for mid-resource languages, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 1703–1714. URL: https://aclanthology.org/2020.acl-main.156. doi:10.18653/v1/2020.acl-main.156.

[9] L. Parisi, S. Francia, P. Magnani, Umberto: An italian language model trained with whole word masking, https://github.com/musixmatchresearch/umberto, 2020.

[10] L. Zhuang, L. Wayne, S. Ya, Z. Jun, A robustly optimized BERT pre-training approach with post-training, in: Proceedings of the 20th Chinese National Conference on Computational Linguistics, Chinese Information Processing Society of China, Huhhot, China, 2021, pp. 1218–1227. URL: https://aclanthology.org/2021.ccl-1.108.

[11] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6. doi:10.18653/v1/2020.emnlp-demos.6.

[12] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations, 2019. URL: https://openreview.net/forum?id=Bkg6RiCqY7.

DH-FBK at HODI: Multi-Task Learning with Classifier Ensemble Agreement, Oversampling and Synthetic Data

Elisa Leonardelli[1], Camilla Casula[1,2]

[1]Fondazione Bruno Kessler, Trento, Italy
[2]University of Trento, Trento, Italy

Abstract

We describe the systems submitted by the DH-FBK team to the HODI shared task, dealing with Homotransphobia detection in Italian tweets (Subtask A) and prediction of the textual spans carrying the homotransphobic content (Explainability - Subtask B). We adopt a multi-task approach, developing a model able to solve both tasks at once and learn from different types of information. In our architecture, we fine-tuned an Italian BERT-model for detecting homotransphobic content as a classification task and, simultaneously, for locating the homotransphobic spans as a sequence labeling task. We also took into account the subjective nature of the task by artificially estimating the level of agreement among the annotators using a 5-classifier ensemble and incorporating this information in the multi-task setup. Moreover, we experimented by extending the initial training data with oversampling (Run 1) and via generation of synthetic data (Run2). Our runs achieve competitive results in both tasks. Finally, we conducted a series of additional experiments and a qualitative error analysis.

Keywords

Multi-task learning, data augmentation, agreement, subjective tasks

***Warning**: This paper contains examples of potentially offensive content.*[1]

1. Introduction

In recent years, social media use has increased globally, with platforms enabling users to post, share and comment about any topic at any time. With the increase of online communication, proliferation of online hateful comments has become a major problem. Natural language processing (NLP) research is essential for the mitigation of online hate speech, as it can help in understanding the phenomenon and assisting in automating the process at a large scale.

The NLP community has been tackling this problem through the creation of datasets and models, especially focusing on some of the most vulnerable communities, such as migrants [2] or women [3]. The application of automatic methods for detecting hate speech targeting LGBTQIA+ people specifically is a recent development, having been addressed for the first time in English and Tamil [4] and more recentely in Locatelli et al. [5].

The evaluation task for Homotransphobia Detection in Italian (HODI) [6] proposed at Evalita 2023 [7], aims to explore homotransphobia on Twitter in Italian, taking a deeper look into an issue that has not been adequately addressed in either the global or Italian NLP communities. To this end, the task organizers released a dataset of 6,000 tweets annotated for homophobic and transphobic content (Subtask A) and highlighting the span range expressing it within the sentence (Subtask B), encouraging the developing of models able to detect homotransphobic content in an accurate and explainable way.

In this paper, we present the submitted systems by the DH-FBK team for the two HODI subtasks. Based on the hypothesis that the two layers of annotations provided are highly correlated and thus knowledge sharing will help with the completion of each task, we implemented a multi-task architecture, similarily to the ones proposed in Ramponi and Leonardelli [8] and Leonardelli and Casula [9]. This setup allows leveraging training signals of related tasks at the same time by exploiting a shared representation in the model. Specifically, we simultaneously train a model on the two HODI subtasks, addressing Subtask A as a classification task, and the extraction of the spans containing homotransphobic language (Subtask B) as a Sequential Labeling (SL) problem, locating the spans by BIO tags [10]. Importantly, this multi-task approach allows to develop a unique model for addressing both tasks, and we are one of the two teams who participated in both tasks. Moreover, given the subjectivity of the task, we add an auxiliary task to the multi-task configuration to incorporate information related to annotator agreement. Previous studies have shown that training on data with low agreement between annotators can lead to a decrease in model performance [11]. However, more recent research has shown that this depends on the source of the disagreement and that the level of agreement should still be taken into account when training

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
eleonardelli@fbk.eu (E. Leonardelli); ccasula@fbk.eu (C. Casula)
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

[1]Profanities have been obfuscated with PrOf (https://github.com/dnozza/profanity-obfuscation) [1]

CEUR-WS.org/Vol-3473/paper31.pdf

[12]. Since disaggregated annotations are not accessible to participants in the task, we estimate agreement levels through the use of an ensemble of 5 classifiers, to imitate annotator judgments, similarly to the work conducted in Leonardelli and Casula [9]. Additionally, following the organizers' suggestion to increase the train size of the data, we experimented with different methods for augmenting the training size, i.e. oversampling [13] and data generation [14].

Our best performing run (Run 1) achieved competitive results, ranking 4th for Subtask A and 2nd for Subtask B.

Finally, we discuss the impact of the different elements we combined in our models by conducting a series of additional experiments in Section 5.2, showing the benefits of augmenting training data, especially using oversampling, and showing the relative beneficial impact of the auxiliary task on agreement, which is effective only in combination with oversampling and not with the synthetic data. We then also conduct a qualitative analysis to discover the most difficult cases.

2. The HODI dataset

The HODI dataset is composed of 6,000 Italian tweets. The tweets have been collected from the 1st of May 2022 to the 31st of August 2022 using a set of 21 keywords associated with language that might potentially target minority groups victims of homotransphobia. Entries are annotated following a two-layer scheme:

1. Homotransphobia detection: a tweet contains homotransphobic language or not (binary).
2. Rationales detection (explainability): when a tweet is considered homotransphobic, the span of text that contains the homotranshphobic part is highlighted (list of character positions).

3. Task description

The organizers provided participants with the HODI dataset, described in Section 2. 5,000 annotated tweets were released during the first phase of the competition, out of which 2,008 labeled as homotransphobic. In a second phase, the remaining 1,000 tweets were released unlabeled as test data. The task is divided into two subtasks, reflecting the layers of annotations of the dataset:

- Subtask A - Homotransphobia detection: binary classification, the goal is to predict whether a tweet is homotransphobic or not.
- Subtask B - Explainability: participants are required to predict the spans of homotransphobic tweet that were responsible for the homotransphobic label of the tweet.

The metric used for evaluation of Subtask A is macro-F_1, while character-based F_1 is used for evaluating Subtask B, similarly to Pavlopoulos et al. [15].

4. Methods

4.1. Multi-task setup

To exploit the strong correlations between the annotations of Subtasks A and B, we used a multi-task learning setup [16], showed in Figure 1. Our model is trained simultaneously on tasks relative to both levels of annotation and, by utilizing a shared representation, all the available information is available to the model. Moreover, the tasks under scrutiny are highly subjective. For instance, we observed some inconsistencies across sentences (for example articles being included/excluded in spans). To leverage the uncertainty around words that are potentially ambiguous, and given that no information about agreement among annotators was released, we 'artificially' created an *agreement* label by using the agreement of an ensemble of 5 classifiers. This procedure is described in more detail in Appendix A. In summary, we use three tasks for our multi-task model: two main tasks corresponding to the two annotation levels (and subtasks) of HODI, and an additional auxiliary task relative to synthetic agreement. The three tasks can be summarized as:

- Homotransphobia (Subtask A): binary classification of homotransphobia.
- Explainability (Subtask B): annotations for this task are released at character level. We convert each sentence from character to word-level annotation, and associate each word to a label for whether it belongs to the homotransphobic span (Explainibility). Moreover, since often spans are comprised of entire phrases, annotations followed sequence labelling, using a BIO tagging scheme [17] in which each word can be at the beginning, inside or outside of a span. After converting the data into this format, Subtask B can be carried out as a sequence labelling prediction task.
- Agreement on Subtask B: it is addressed as sequence labelling at word-level. It can assume values between [0-5], reflecting how many of the 5-classifier ensemble, described in Appendix A, predict a specific word in agreement with the gold label of Subtask B.

4.2. Synthetic Data

The use of synthetic data has been proposed as a method to increase the amount of available training data for hate speech and offensive language detection tasks, especially

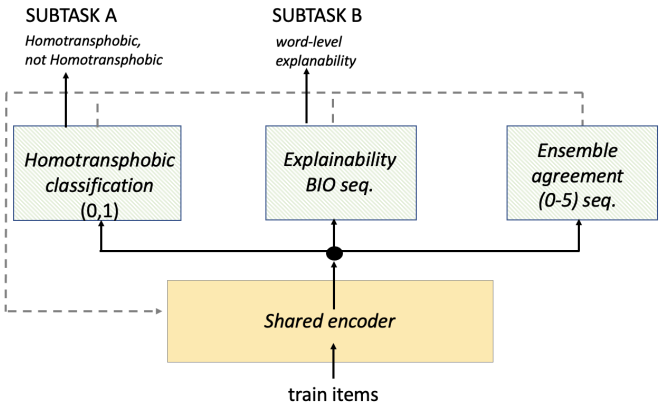


Figure 1: The multi-task configuration of the model we used for Subtask A and B predictions.

when relying on machine-generated data [18, 19]. Although data augmentation using generative models has been found to not always be reliable in improving models [14], we aim at exploring whether it can help the performance of our models for the HODI task.

A widely used method of generating synthetic data consists in fine-tuning a generative model on annotated data and then using it for generating new sequences. These generated sequences are then passed through a classifier in order to confirm the label assignment made by the generator, since generative models are not always reliable in their label assignment [20].

While the majority of works that exploit model-generated data for the detection of offensive language have no particular focus on any target category or phenomenon, our experiments are focused on specifically detecting homotransphobia. Because of this, the generated texts should be correct regarding both the label and the focus on the phenomenon. In part due to this, and in part to the limited availability of generative large language models for Italian, we decide to generate new data using an encoder-decoder model trained on Italian, IT5 [21], in its 738M-parameter (`large`) configuration. The details of our data augmentation process can be found in Appendix B.

Given that the augmentation process provides us with synthetic examples annotated for Subtask A (Homotransphobia detection), but not for Subtask B (detection of rationales), we additionally estimate Subtask B labels for the generated instances, using the model of the first submitted run (generated data were used only in run2), while the agreement for the auxiliary task was estimated using the ensemble classifier described in Appendix A.

4.3. Experimental Setup

The models developed for the two runs submitted by our team, are both based on a pre-trained Italian BERT [2]. For fine-tuning the models in the multi-task setup described in Figure 1, we employed the MaChAmp v0.2 toolkit [22], a tool that supports a variety of standard NLP tasks out-of-the-box, also in a multi-task setup. We employed the pre-trained BERT as our shared encoder for all tasks, while a separate decoder is utilized by each task. We fine-tuned the model (110M parameters) for 15 epochs on a single GPU [3], using default MaChAmp hyperparameters[4]. For the training process, we assign each class equal weight to guarantee minority classes are not underrepresented. We introduced also loss weights for the multi-task learning loss, calculated as $L = \sum_t \lambda_t L_t$, where L_t accounts for the loss for task t and λ_t being the respective weighting parameter. We set $\lambda_t = 0.8$ for the primary tasks, and $\lambda_t = 0.5$ for the auxiliary tasks.

4.4. Submitted Systems Description

For the competition, we submitted two different runs with predictions by models created using the same setup described in Section 4.3 and Figure 1, but trained on different sets of data. Starting from the suggestion from the organizers to augment the size of the training set, we experimented with oversampling and data generation in the following way:

- **Run 1:** the data made available from the organizers are oversampled by repeating them twice.

[2] dbmdz/bert-base-italian-cased
[3] NVIDIA Titan Xp
[4] Defaults MaChAmp hyperparameter settings used for all our experiments; Optimizer: AdamW; β_1, β_2 = 0.9; Dropout = 0.2; Batch Size = 32; Learning rate = 0.0001

- **Run 2:** In addition to oversampling the HODI data, similarly to run1, we add 4,000 synthethically generated examples (see Section 4.2).

We split HODI data into 90% training set and 10% development set. For Run2, the synthetic examples were added to the training set.

5. Evaluation

5.1. Results

Table 1 shows the official results of our submissions for Subtasks A and B. All runs for both tasks beat the organizers' baseline.

For Subtask A we report macro-averaged F_1 score and overall rank of our runs, as well as those of the teams who performed better than us and the baseline. Our best performance (Run 1) obtained a macro F_1 score of 0.795, ranking 4th out of 18 submitted runs (3rd out of 8 teams), while run 2 ranked 7th out of 18 submissions (4th out of 8 teams).

For Subtask B we report the overall ranking, given that the leaderboard is short and only another team participated in Subtask B. One run of the other team participating in this task beat our result, while our best scoring run (Run 1) ranked 2nd.

Table 1
Overview of the results

Subtask A	Run	F1	Team
1	Run 3	0.8108	metzi
2	Run 2	0.8	metzi
3	Run 1	0.7959	odang4hodi
4	**Run 1**	**0.7950**	**DH-FBK**
5	Run 2	0.7942	extremITA
6	Run 2	0.7920	odang4hodi
7	**Run 2**	**0.7837**	**DH-FBK**
		...	
13	-	0.6691	baseline_a
19		...	
Subtask B			
1	Run 2	0.723	extremITA
2	**Run 1**	**0.705**	**DH-FBK**
3	**Run 2**	**0.701**	**DH-FBK**
4	Run 1	0.66	extremITA
5	-	0.205	baseline_b

5.2. Additional experiments

Regarding the impact of generated data, when adding the synthetic data in the training (Run 1) performance decreases in both tasks, showing that the augmentation with generated data does not improve the generalization of models compared to oversampling. In fact, we hypothesize that the addition of synthetic data might push models to be over-reliant on specific identity terms or profanities, hurting its generalization capabilities, a phenomenon that has been observed in data augmentation using generative models [14]. Moreover, to dissect the impact of oversampling and the impact of the auxiliary task, we run a series of additional experiments[5]. Results are shown in Table 2. To evaluate the role of oversampling we replicate the setup of the two submitted runs but omitting the oversampling of the HODI data from the training (Exp 1 and Exp 2). By comparing results (Run 1 vs Exp 1 and Run 2 vs Exp 2), we can observe how oversampling the data is generally beneficial, especially if no synthetic data are used. Moreover, in Exp 3 and Exp 4 we replicate the submitted experiments but exclude the auxiliary task. By comparing Run1 and Exp 3, we can observe how in this case the task is indeed beneficial, while it is not when comparing Run 2 and Exp 4, where synthetic examples are part of the training data. This suggests that the estimation of agreement for generated data might not be informative.

5.3. Qualitative Error Analysis - Subtask A

To perform a qualitative analysis on the most problematic tweets, we isolated the tweets that were incorrectly classified by all models in Table 2. The most consistent false negative regards the missed detection of tweets containing a specific offensive slang word (*f*mminiello*). One possible reason is that this word is not generally common (as it belongs to a local language variety), and it was not present in the training set. Observing the posts incorrectly classified as homotransphobic by the models, we identified (doubtful) sense of humour or metaphorical expressions (*andare a fare in culo, essere fr*cio col culo degli altri*) as possible reasons. Another possible reason could also be over-reliance on specific terms.

6. Conclusions

We described our participation to the HODI evaluation task at Evalita 2023. We used a multi-task learning approach to share representations between the two tasks involved and, additionally, considering the subjectivity of the task, we incorporated inter-annotator agreement information into our framework, estimating them with a 5-classifier ensemble. We experimented augmentation of the training data available by oversampling them and via generated data. We were one of the few teams who participated in both tasks, and our systems performed competitively.

Moreover, we conducted an analysis on the role and impact of the various aspects we combined. Our results show oversampling is generally beneficial, especially

[5]The organizers released the labels for the test set after the closing of the evaluation phase

Table 2
Classifier performance by varying the subsets of the training set and the auxiliary task

Exp.	Multi-task setup		Training		F_1 subtasks	
	main	auxiliary	on	size	A	B
Run 1 (sub)	subtask A & B	ens.-agr.	2 x data	10,000	**0.795**	**0.705**
Run 2 (sub)	subtask A & B	ens.-agr.	2 x data + synth. data	14,000	0.784	0.701
Exp 1 (post)	subtask A & B	ens.-agr.	data	5,000	0.787	0.692
Exp 2 (post)	subtask A & B	ens.-agr.	data + synth. data	9,000	0.789	0.691
Exp 3 (post)	subtask A & B	-	2 x data	10,000	0.785	0.694
Exp 4 (post)	subtask A & B	-	2 x data + synth. data	14,000	0.788	0.698

when combined with the auxiliary task on agreement. The usage of generated data instead has limited benefits, compared to oversampling or additional auxiliary tasks. Finally, performing a qualitative analysis on the most frequent causes of error, we identified specific homotransphobic slang terms that were problematic to be identified by our models.

Acknowledgments

The work of Elisa Leonardelli was partially funded by the StandByMe European project (REC-RDAP-GBV-AG-2020) on "Stop online violence against women and girls by changing attitudes and behaviour of young people through human rights education" (GA 101005641). Her research was also supported by the StandByMe 2.0 project (CERV-2021-DAPHNE) on "Stop gender-based violence by addressing masculinities and changing behaviour of young people through human rights education" (GA 101049386).

References

[1] D. Nozza, D. Hovy, The state of profanity obfuscation in natural language processing scientific publications, in: Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, 2023.

[2] T. Bourgeade, A. T. Cignarella, S. Frenda, M. Laurent, W. Schmeisser-Nieto, F. Benamara, C. Bosco, V. Moriceau, V. Patti, M. Taulé, A multilingual dataset of racial stereotypes in social media conversational threads, in: Findings of the Association for Computational Linguistics: EACL 2023, 2023, pp. 674–684.

[3] E. Fersini, P. Rosso, M. Anzovino, Overview of the task on automatic misogyny identification at ibereval 2018., Ibereval@ sepln 2150 (2018) 214–228.

[4] B. R. Chakravarthi, R. Priyadharshini, R. Ponnusamy, P. K. Kumaresan, K. Sampath, D. Thenmozhi, S. Thangasamy, R. Nallathambi, J. P. McCrae, Dataset for identification of homophobia and transophobia in multilingual youtube comments, arXiv preprint arXiv:2109.00227 (2021).

[5] D. Locatelli, G. Damo, D. Nozza, A cross-lingual study of homotransphobia on twitter, in: Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP), 2023, pp. 16–24.

[6] D. Nozza, A. T. Cignarella, G. Damo, T. Caselli, V. Patti, HODI at EVALITA 2023: Overview of the Homotransphobia Detection in Italian Task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[7] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[8] A. Ramponi, E. Leonardelli, Dh-fbk at semeval-2022 task 4: leveraging annotators' disagreement and multiple data views for patronizing language detection, in: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), 2022, pp. 324–334.

[9] E. Leonardelli, C. Casula, Dh-fbk at semeval-2023 task 10: Multi-task learning with classifier ensemble agreement for sexism detection, in: Proceedings of the 17th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1894–1905. URL: https://aclanthology.org/2023.semeval-1.261.

[10] Q. Zhu, Z. Lin, Y. Zhang, J. Sun, X. Li, Q. Lin, Y. Dang, R. Xu, Hitsz-hlt at semeval-2021 task 5: Ensemble sequence labeling and span boundary detection for toxic span detection, in: Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021), 2021, pp. 521–526.

[11] E. Leonardelli, S. Menini, A. Palmero Aprosio, M. Guerini, S. Tonelli, Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 10528–10539. URL: https://aclanthology.org/2021.emnlp-main.822. doi:10.18653/v1/2021.emnlp-main.822.

[12] M. Sandri, E. Leonardelli, S. Tonelli, E. Jezek, Why don't you do it right? analysing annotators' disagreement in subjective tasks, in: Proceedings of the 2023 Conference of the European Chapter of the Association for Computational Linguistics, 2023.

[13] E. Leonardelli, S. Menini, S. Tonelli, Dh-fbk@ haspeede2: Italian hate speech detection via self-training and oversampling., in: Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), volume 2765, 2020.

[14] C. Casula, S. Tonelli, Generation-based data augmentation for offensive language detection: Is it worth it?, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 3359–3377. URL: https://aclanthology.org/2023.eacl-main.244.

[15] J. Pavlopoulos, J. Sorensen, L. Laugier, I. Androutsopoulos, Semeval-2021 task 5: Toxic spans detection, in: Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021), 2021, pp. 59–69.

[16] R. Caruana, Multitask learning, Machine learning 28 (1997) 41–75.

[17] J. Lafferty, A. McCallum, F. C. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001).

[18] T. Wullach, A. Adler, E. Minkov, Fight Fire with Fire: Fine-tuning Hate Detectors using Large Samples of Generated Hate Speech, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 4699–4705. URL: https://aclanthology.org/2021.findings-emnlp.402.

[19] T. Hartvigsen, S. Gabriel, H. Palangi, M. Sap, D. Ray, E. Kamar, ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 3309–3326. URL: https://aclanthology.org/2022.acl-long.234. doi:10.18653/v1/2022.acl-long.234.

[20] A. Anaby-Tavor, B. Carmeli, E. Goldbraich, A. Kantor, G. Kour, S. Shlomov, N. Tepper, N. Zwerdling, Do Not Have Enough Data? Deep Learning to the Rescue!, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 7383–7390. doi:10.1609/aaai.v34i05.6233.

[21] G. Sarti, M. Nissim, IT5: Large-scale text-to-text pretraining for italian language understanding and generation, ArXiv preprint 2203.03759 (2022). URL: https://arxiv.org/abs/2203.03759.

[22] R. van der Goot, A. Üstün, A. Ramponi, I. Sharaf, B. Plank, Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Online, 2021, pp. 176–197. URL: https://aclanthology.org/2021.eacl-demos.22. doi:10.18653/v1/2021.eacl-demos.22.

Appendix

A. Ensemble agreement

For posts (of the training set) that were annotated as homotransphobic, we aim at obtaining an approximation of the agreement level on each word of the post, as being considered part of the span is correlated with labeling the post as homotranspophobic. This information is then exploited as additional information in our multi-task training setup, specifically as an extension to the sequence labelling prediction of Subtask B.

We split the training data X provided by the HODI organizers in 5 folds $X_1, X_2, ..., X_5$, creating 5 separate train/validation splits, being careful that each item of the train appears in the validation set of one fold. We employ an ensemble of classifiers, a method first suggested by Leonardelli et al. 2021 [11], where each classifier of the ensemble is trained using slightly different configurations by varying the initial conditions such as the initial seed and the number of epochs, so that the 5 classifiers produce similar but not identical predictions. The classifiers are produced in the multi-task setup showed in Figure 1, but without the Auxiliary Task on agreement. In this manner, we have ensemble predictions for each of the entries of the training data. Based on the predictions of the classifiers, we assign ensemble agreement labels to the validation set (at a word-level) of the current fold based on how many classifiers agree with the actual gold annotation. The ensemble agreement label is thus a number between 0 and 5. We consider this information as proxy for item's difficulty and annotators' disagreement.

B. Data augmentation pipeline

The pipeline we follow for augmenting the available data for the task is as follows:

1. We fine-tune a classifier (in our case a BERT-base model trained on Italian) [2] on the HODI training data.
2. We fine-tune IT5-Large on the same training data, formatting the task so that the input is '*Scrivi un tweet:*' or ´*Scrivi un tweet omotransfobico:*' ('*Write a tweet:*' or ´*Write a homotransphobic tweet:*') depending on the gold label of each example, and the output is the actual post.
3. We use the fine-tuned IT5 model to generate new data, using the same type of input we use in Step 2.
4. We filter the generated data using the fine-tuned classifier from Step 1, keeping only the examples for which the label assignment is the same for the classifier and the generator [20, 14]. We additionally remove duplicates and normalize URLs as URL.
5. We rank generated examples based on the confidence of the classification model we used for filtering, retaining the top 2,000 examples for each class. This number is chosen in order to ideally double the size of the dataset, and we use generated examples that are equally split among the labels so as to artificially mitigate the class imbalance.

MULTI-Fake-DetectiVE at EVALITA 2023: Overview of the MULTImodal Fake News Detection and VErification Task

Alessandro Bondielli[1,2,*], Pietro Dell'Oglio[3,4], Alessandro Lenci[2], Francesco Marcelloni[2], Lucia C. Passaro[2] and Marco Sabbatini[2]

[1]Department of Computer Science, University of Pisa, Italy
[2]Department of Philology, Literature and Linguistics, University of Pisa, Italy
[3]Department of Information Engineering, University of Florence, Italy
[3]Department of Information Engineering, University of Pisa, Italy

Abstract

This paper introduces the MULTI-Fake-DetectiVE shared task for the EVALITA 2023 campaign. The task was aimed at exploring multimodality within the realm of fake news and intended to address the problem from two perspectives, represented by the two sub-tasks. In sub-task 1, we aimed to evaluate the effectiveness of multimodal fake news detection systems. In sub-task 2, we sought to gain insights into the interplay between text and images, specifically how they mutually influence the interpretation of content in the context of distinguishing between fake and real news. Both perspectives were framed as classification problems.

The paper presents an overview of the task. In particular, we detail the key aspects of the task, including the creation of a new dataset for fake news detection in Italian, the evaluation methodology and criteria, the participant systems, and their results. In light of the obtained results, we argue that the problem is still open and propose some future directions.

Keywords

Fake News, Fake news detection, Multi-modality, Vision-Language models, Large Language Models

1. Introduction and Motivation

Recent years have seen a great increase in the online proliferation of disinformation and fake news [1]. This is especially true in the context of real-world events that are reported as breaking news. It is often the case that entities with malicious intents exploit breaking news to push their own agenda by distorting facts and intentionally publishing false or misleading information.

Distorted uses of online social media have been made mostly evident in the last few years by the first so-called *infodemic* following the COVID-19 pandemic [2], in what has been defined by several authors as a Post-Truth Era [3] dominated by emotions and pseudo-facts [4]. This phenomenon has grown further with the outbreak of the Russian war against Ukraine. Like in all conflicts, disinformation has become a powerful strategic weapon.

These issues have led over the years to the creation of numerous initiatives for independent fact-checking and fake news detection, and the topic has increased its relevance in the research community. The literature on issues related to fake news detection, disinformation, and fact-checking, is constantly growing despite the inherent challenges and many facets of the problem.

A large number of approaches and techniques have been proposed for content verification and fake news detection in a uni-modal setting. Most of the proposed approaches use either the actual content of the news (i.e., the text itself), its context (e.g., social network structures, temporal information), or a combination of both [5]. Most modern systems typically leverage transformer models with additional information [6].

It is clear that the easiest way to spread disinformation is in textual form. However, online outlets and social media allow for other modalities as well. Images for example can be leveraged in the context of disinformation and fake news in different ways: first, the inclusion of images in malicious content can be leveraged as a way to provide more credibility for the text containing the fake news; second, images could be described in such ways that their original content is misinterpreted by readers, leading again to disinformation; finally, they can be used in an attempt to increase the post attraction and get the fake news shared by as many social media users as possible.

We can argue that multimodal scenarios may be considered as closer in nature to real-world ones examin-

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
*Corresponding author.
alessandro.bondielli@unipi.it (A. Bondielli); pietro.delloglio@unifi.it (P. Dell'Oglio); alessandro.lenci@unipi.it (A. Lenci); francesco.marcelloni@unipi.it (F. Marcelloni); lucia.passaro@unipi.it (L. C. Passaro); marco.sabbatini@unipi.it (M. Sabbatini)
0000-0003-3426-6643 (A. Bondielli); 0000-0002-0793-5226 (P. Dell'Oglio); 0000-0001-5790-4308 (A. Lenci); 0000-0002-5895-876X (F. Marcelloni); 0000-0003-4934-5344 (L. C. Passaro); 0000-0002-8837-6592 (M. Sabbatini)
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper32.pdf

ing social media data. Nevertheless, multimodality has received relatively less attention over the years in this context [4]. This is rapidly changing, with a number of international multimodal shared tasks being organised for fake news and propaganda detection, fact-checking and related areas [7, 8, 9, 10]. Nevertheless, models combining multiple modalities for detecting fake news remain a major open challenge in the literature, as well as datasets including different modalities and different sources of fake news [4]. Moreover, we believe that a fundamental step towards a more nuanced understanding of the problem lies in actually understanding and modelling the interplay between the different modalities in generating disinformation.

In this context, we propose MULTI-Fake-DetectiVE[1] as part of the EVALITA 2023 Evaluation campaign [11]. The task is aimed at addressing both the textual and visual aspects of fake news on social media and online news outlets, from two key perspectives: we want to model fake news detection from a multimodal perspective, and we are interested in exploring how images and texts interact and influence each other in the context of real and fake news. Further, we contribute to this research area by creating a dataset of social media posts from Twitter and news articles regarding the Russian-Ukrainian war including fake and real news.

2. Definition of the task

MULTI-Fake-DetectiVE includes two sub-tasks. Both are formulated as multi-class classification problems. In the first sub-task, given a piece of content (i.e., a social media post or a news article) that includes both a visual and a textual component, the goal is to determine its likelihood of being a *real* or a *fake* news. In the second sub-task, given a text and an accompanying image, the goal is to decide whether their combination is aimed at *misleading* the interpretation of the reader about one or the other, or not. Note that for both sub-tasks we consider the visual component as all the images provided with a given textual content (i.e., news article or social media post). Thus, for example, if a tweet includes three images, and one of them is misleading, the expected label will be *misleading*.

In the following, we describe in detail both sub-tasks.

2.1. Sub-task 1: Multimodal Fake News Detection

The first sub-task is structured as a multi-class classification problem in a multimodal setting. The problem is defined as follows: given a piece of content $c = \langle t, v \rangle$ which includes a textual component t and a visual component v (i.e., one or more images), classify it into one of the following classes:

Certainly Fake: news that is certain to be fake, whatever the context.

Probably Fake: news that is likely to be fake, but may include some real information or at the very least be somewhat credible.

Probably Real: news that is very credible but still retains some degree of uncertainty about the provided information.

Certainly Real: news that is certain to be real and incontestable, whatever the context.

The classes refer to the informational content as a whole, and not to its single components. For example, a fake piece of news including a real image (e.g., in a misleading context) is still probably (or certainly) fake.

2.2. Sub-task 2: Cross-modal relations in Fake and Real News

The second sub-task is aimed at assessing how the two modalities (i.e., textual and visual) interact in the context of fake and real news. Our goal is to understand how images and texts in fake and real news can lead to misleading interpretations of the content pertaining to the other modality and to the whole news.

The sub-task is a three-class classification problem, and is defined as follows: given a piece of content $c = \langle t, v \rangle$ which includes a textual component t and a visual component v, decide whether their combination is:

Misleading: one between the textual and visual components is used deceptively to lead to misinterpretation of the other.

Not Misleading: the combination of the visual and textual component does NOT lead to misinterpretation of the news.

Unrelated: the visual component is not related to the text component, or does not add information to the text component or does not change its interpretation in a meaningful way.

3. Dataset

The dataset for the shared task includes social media posts and news articles, containing both a textual and a visual component, concerning one or more real world events that are known to have been subject to the generation of fake news. In particular, the dataset focuses on

[1] https://sites.google.com/unipi.it/multi-fake-detective/home

the Ukrainian-Russian war, and includes data in a time span going from February 2022 to December 2022.

The dataset is composed of two sub-datasets, one for each sub-task. Each is further split into a training set and two different test sets. More specifically, the dataset for each sub task is divided as follows:

Training Set: the training data provided to participants. It includes data from February 2022 to September 2022.

Test Set (Official): the official test set used for evaluation. It includes data from October 2022 to December 2022.

Test Set (Additional): an additional batch of test data including data from the same time window as the training set.

The Official Test Set was developed to challenge participating systems to classify fake news and misleading content in a more real-world scenario (i.e., different time windows that might determine different data distributions). The Additional Test Set was instead aimed at giving us a clearer picture of how participating systems are resilient to changes in the context over time [12]. Note that the evaluation on the Additional Test Set was not mandatory.

The dataset is available for download on the website of the task.[2]

3.1. Data Collection and annotation

The dataset was collected and annotated via crowdsourcing following a multi-step process heavily inspired by the one proposed in [13]. First, we broadly collected Twitter data regarding the Ukrainian-Russian war in the chosen time span. To collect such data, we chose a set of keywords representative of the conflict, e.g., "Ucraina, Russia, Putin, Zelensky". In addition to this, we collected texts and images for news articles that were in the tweets. At this stage, the data were collected regardless of the sub-tasks.

Then, we exploited a manually collected set of verified fake news and misleading claims (henceforth referred to as *seed fake news and misleading claims*) to generate the dataset for each sub-task. We took into account different news outlets reporting on the fake news and independent fact-checking websites. These seed fake news and misleading claims were intended to serve a dual purpose. On the one hand, we used them to filter the original dataset by considering their similarity with data samples. This was done to ensure that: i) the resulting datasets would include only relevant elements (i.e., that actually refer to the Ukrainian-Russian war), and ii) the class distribution for both sub-tasks was not too skewed in favour of real news and not misleading claims, as it would have been in an uncontrolled scenario. On the other hand, the seed fake news and misleading claims served as context for the annotation process. Specifically, we used Prolific[3] to obtain labels for our dataset. For each sub-task, we provided annotators with the seed fake news and misleading claims as context, and asked them to label a few of the data samples. Each data sample was labelled by at least five different annotators. We collected human annotations and kept only data samples for which at least 3 out of the 5 annotators provided the same label.

Table 1
Dataset size for sub-task 1.

	C.F.	P.F.	P.R.	C.R.
Train	153	219	476	199
Test (official)	16	52	106	21
Test (additional)	27	58	101	32
Total	196	329	683	252

Datasets sizes and class distributions are reported in Tables 1 and 2 for, respectively, sub-tasks 1 and 2. In sub-task 1, inter-annotator agreement was calculated as the average Spearman correlation coefficient between annotator pairs, considering the ordered nature of the labels. The average correlation was 0.43 (σ = 0.04). In sub-task 2 we employed Fleiss' Kappa to measure the inter-annotator agreement since the labels were not inherently ordered. We obtained k = 0.25.

Participants were provided with a TSV file containing IDs, URLs, and numeric labels representing classes for sub-task 1 and sub-task 2. The label was excluded from the test set during the evaluation period. Participants had the option to download the data using their preferred method or utilise a provided download script. The script offered participants access to textual data, including metadata such as URLs, data type (e.g., tweet or article), and creation date if available, as well as associated images. Authorship information was not provided with the data. Note that while the datasets were treated separately for annotation, some data samples could be present in both sub-tasks. In such cases, the ID associated with the data point remained consistent across the two sub-tasks.

3.2. Copyright and Content Warning

The dataset includes tweets and news articles. The provided download script performs a coarse-grained anonymization of the data (e.g. by removing author information).

Upon download, users agree not to share the material they receive both during and after the competition. The

[2] https://sites.google.com/unipi.it/multi-fake-detective/data

[3] prolific.co

Table 2
Dataset size for sub-task 2.

	Misl.	Not Misl.	Unrel.
Train	373	546	417
Test (official)	45	75	99
Test (additional)	67	84	89
Total	**485**	**705**	**605**

data for the MULTI-Fake-Detective tasks is to be used for research purposes only. Note that by receiving the data users implicitly agree to Twitter Terms of Service, Privacy Policy, Developer Agreement, and Developer Policy for academic researchers.

We do not share responsibility for the contents of the dataset. Downloaded texts and images may include copyrighted material and sensitive contents. The downloaded data and the provided labels do not reflect in any way the social and political views of the task organisers.

4. Evaluation measures

Participants were allowed to present up to four different systems for predicting labels on the official test set, with one system marked as *primary*. Results for primary systems were used as basis for the final ranking. Specifically, the ranking was calculated based on the **weighted average F1-score** of the systems. The same evaluation procedure and criteria was applied to both sub-tasks. The evaluation procedure was conducted by means of an evaluation script (available to participants).

Note that due to restrictions in data distribution (see Section 3), not all participants may have had access to the exact same test dataset. For example, articles/tweets in the dataset may have been removed by the authors during the evaluation window. To ensure fair competition, we evaluated and ranked the systems only on the subsets of the test sets for which all the participants were able to provide a label.

4.1. Baseline models

Participating systems were evaluated against each other and against a set of baseline models.

Specifically, we proposed two different classification models, namely a Support Vector Machine (SVM) and a Multi-Layer Perceptron (MLP), with three different feature sets as the baseline models. As for the feature sets, we considered:

Text-only features extracted with a multilingual BERT model [14].

Image-only features extracted with ResNet-18 [15].

Multimodal features obtained by concatenating the text-only and image-only features.

All models were trained using the default parameters from scikit-learn[4].

To ensure fair reproducibility and comparisons, the baseline models and the evaluation scripts are available on the website of the task. [5]

5. Participating systems and results

A total of four teams participated in MULTI-Fake-DetectiVE. All four teams participated to sub-task 1 (Multimodal Fake News Detection), and two of them also participated to sub-task 2 (Cross-modal relations in Fake and Real News). The proposed approaches are quite different. We can distinguish between two truly multimodal approaches and two text-oriented ones. In the following, we broadly describe the core systems of participating teams.

PoliTo [16] participated to both sub-tasks with an approach focused on refining FND-CLIP [17], a fake news detection multimodal model based on CLIP [18]. Authors proposed several refinements to the original model via ad-hoc extensions including sentiment-based text encoding, image transformations in the frequency domain, and data augmentation via back translation. The final model for both sub-tasks is an ensemble that combines predictions for all the extensions.

AIMH [19] participated to both sub-tasks with a vision-text dual encoder approach. They used ViT to encode images and RoBERTa/BERT to encode texts. Authors experimented with different inputs for their model. They generated image captions and automatically translated Italian texts to English. They tested various input combinations and chose to use English texts and images as inputs for their final model.

ExtremITA [20] participated with a text-only approach aimed at solving all EVALITA tasks via prompt engineering of Large Language Models. The team proposed two Italian models, an encoder-decoder based on T5 [21] and an instruction-tuned decoder-only model based on LLaMA [22].

HIJLI-JU-CLEF [23] proposed a text-oriented model to solve sub-task 1. The model uses a pre-trained

[4] https://scikit-learn.org/stable/supervised_learning.html
[5] https://sites.google.com/unipi.it/multi-fake-detective/tasks-and-evaluation

Table 3
Sub-task 1 - Official Test Set.

Rank	TEAM-RUN	F1-Score
1	**Polito-P1**	**0.512**
2	extremITA-camoscio_lora	0.507
3	AIMH-MYPRIMARYRUN	0.488
4	*Baseline-SVM_TEXT*	0.479
5	*Baseline-SVM_MULTI*	0.463
6	*Baseline-MLP_TEXT*	0.448
7	*Baseline-MLP_IMAGE*	0.402
8	HIJLI-JU-CLEF-Multi	0.393
9	*Baseline-SVM_IMAGE*	0.386
10	*Baseline-MLP_MULTI*	0.374

Table 4
Sub-task 1 - Additional Test Set.

Rank	TEAM-RUN	F1-Score
1	**extremITA-camoscio_lora**	**0.464**
2	PoliTo	0.460
3	extremITA-it5	0.348

model applied to image captions and a pre-trained model to translate Italian texts into English. Finally, it uses translated texts and captions as inputs to either a BiLSTM or a Transformer-based model to classify the data.

Participating systems and baselines were evaluated and ranked according to the evaluation criteria described in Section 4. In the following, we present the results on each sub-task.

5.1. Results of Sub-task 1

All participating systems attempted to solve the Multimodal Fake News Detection sub-task. Tables 3 and 4 detail the obtained results of each system including baselines on the Official and Additional test sets, respectively.

As for the Official test set, the PoliTo ensemble model and the LLaMA-based ExtremITA model ranked first and second, with close results. The AIMH vision-text dual encoder model ranked third. All three models were able to outperform all baseline systems, albeit marginally. The best performing baselines are the text-based and multimodal SVM models. The other baseline models performed significantly worse. Finally, the HIJLI-JU-CLEF text-oriented system was able to outperform two out of the six baseline models proposed, ranking fourth among participants and eighth globally.

As for the Additional test set, the best performing model was the LLaMA-based ExtremITA, closely followed by the PoliTo ensemble approach. The T5-based ExtremITA model performed significantly worse. The AIMH and HIJLI-JU-CLEF systems did not participate in the additional evaluation.

Table 5
Sub-task 2 - Official Test Set.

Rank	TEAM-RUN	F1-Score
1	**Polito-P1**	**0.517**
2	*Baseline-MLP-TEXT*	0.506
3	*Baseline-SVM-TEXT*	0.482
4	*Baseline-MLP-MULTI*	0.461
5	*Baseline-SVM-MULTI*	0.442
6	*Baseline-SVM-IMAGE*	0.436
7	AIMH-MYPRIMARYSUB	0.421
8	*Baseline-MLP-IMAGE*	0.373

5.2. Results of Sub-task 2

Only the truly multimodal models participated in the Cross-modal relations in Fake and Real News sub-task. This is due to the fact that the task is inherently multimodal, and cannot be modelled properly with text-only models: the relationship between image and text features lies at the core of the task, and thus images have to be modelled in some capacity to face it effectively.

Table 5 shows the results obtained by each system, including baselines, on the Official test set. The PoliTo ensemble model ranked first, outperforming all baseline models, while the AIMH vision-text dual encoder model outperformed only the image-only MLP model, ranking seventh. Among baselines, surprisingly the best-performing ones are the text-only models, followed by the multimodal ones. We suspect that the text-only baseline performances are to be attributed to chance rather than to their effective modeling of the problem.

Only the PoliTo team participated in the Additional evaluation, obtaining a weighted average F1-Score of 0.61.

6. Discussion

We can draw some interesting insights from comparing the different proposed models both in terms of their architectures and obtained results.

General findings. First, we can argue that multimodal fake news detection and cross-modal analysis of images and texts in the context of fake news are two rather challenging tasks. As shown by agreement metrics, it was a challenging task for annotators as well (see Sec. 3). This is reflected also by the fact that even the best performing systems were not able to considerably improve over the baseline models results.

As for performances between the Official and Additional test sets, we saw a rather large discrepancy among tasks. We expected systems to perform better on data from the same time period. This appears to be true for sub-task 2, but not for sub-task 1. Note however that the only true comparison can be made on the PoliTo system, as it is the only one that participated to the Official and Additional evaluation for both tasks.

Finally, we must point out the weaknesses of the baseline models. While participating systems were able to perform consistently across sub-task and test sets, the performances of the baseline systems exhibit significant variability, with the relative rankings and performance disparities among models varying across tasks. This suggests that the baselines are unable to adequately model features of both modalities and to leverage them for the tasks.

The architectures of the systems. In sub-task 1, only two out of the four participating systems could be considered as truly multimodal, since they explicitly model image-level features (i.e., with an image encoder model). They are quite similar in principle: both leverage a dual encoder architecture [24] with a classification layer. A ViT image encoder was chosen by both approaches, albeit trained on different data. The text encoders employed in the AIMH system are RoBERTa (for English translations) and an Italian version of BERT for original texts. The PoliTo team uses the FND-CLIP text encoder, which is based on GPT instead. The popularity of CLIP-like Vision-Language models is evident due to their versatility and ease of adaptation for various scenarios and downstream tasks, including fake news detection.The main differences are the extensions (e.g., the inclusion of sentiment-aware text features and image transformations) proposed by the PoliTo team. Authors report performance increases with all proposed extensions, with the ensemble classifier performing best. The remaining two systems either disregard images due to the model architecture (ExtremITA) or consider their automatically generated caption (HIJLI-JU-CLEF), shifting the problem to a text-only space.

The role of textual content. The results of sub-task 1 do not directly highlight a clear advantage of a multimodal approach over a uni-modal one. Two out of the three models which outperform all the baselines are actually multimodal, with the PoliTo FND-CLIP-IT ensemble outperforming all the others. However, the runner-up was the text-only Italian LoRA based on LLaMA with near identical performances.

We can hypothesise that while modelling images in conjunction with text is actually helpful for determining whether a piece of content is a real or a fake news, a large part of the key information to answer the question lies within the textual content. If we assume this, it is easier to understand how model scale also plays a crucial role in performances. The only Large Langauge Model (LLM) presented can get close to the performances of a more refined and nuanced approach in a few-shot setting via prompting. Note that this may hold true regardless of the fact that Italian pre-trained LLMs are still not consistently outperforming all other approaches as their English counterparts due to their sheer size [25].

The importance of additional processing. Sub-task 2 was specifically developed to frame the problem as a multimodal one. Only the two truly multimodal systems participated. As previously discussed, both systems employ a similar architecture and are arguably comparable in terms of model sizes. Thus, we can hypothesize that the difference in performances both in sub-task 1 and 2 may be attributed mostly to the additional processing and extensions applied by the PoliTo system. We could further argue that the tasks are both very complex and nuanced, and additional forms of processing and/or features may provide important benefits in this scenario, rather than sole reliance on textual and visual features extracted from pre-trained models.

7. Conclusions and future directions

In this paper, we presented the MULTI-Fake-DetectiVE shared task for EVALITA 2023. The task was focused on multimodality in the context of fake news. We considered the problem from two perspectives: we wanted to assess fake news detection systems in a multimodal setting, and we wanted to understand how text and images influence each other in the interpretation of a piece of content in the context of fake and real news. We framed both as classification problems.

We saw an interesting degree of variety among proposed systems, which we categorized as truly multimodal or text-oriented. By analyzing the proposed approaches and their results, we can summarise our findings as follows. First, multimodal fake news detection is a very challenging task, especially when considering near real-world scenarios. Second, we saw that for similar vision-language models, both in terms of architecture and model scale, extending the boundaries of the problem by considering additional and/or alternative processing strategies, including affective-oriented features and image processing in the frequency domain, is highly beneficial. Third, we saw that model scale plays also an important role, with pre-trained LLMs approaching the performances of thoroughly fine-tuned systems.

Our findings suggest that the problem is still open and that moving forward it could be advantageous to jointly leverage the advantages of the best-performing approaches, for instance by focusing on Large pre-trained Vision-Language models augmented with additional features (e.g., by using available emotive resources [26]) via either fine-tuning or appropriate prompt tuning/engineering.

Due to the current challenges posed by deceptive or misleading content on social media, we believe that an effective understanding and modelling of such a complex problem may prove to be highly beneficial in contrasting online disinformation. In this regard, the MULTI-Fake-DetectiVE task, including the proposed approaches and the provided datasets, may serve the Italian NLP community as an initial stepping stone in addressing this issue for the Italian language.

Acknowledgments

This research was partially supported by the Italian Ministry of University and Research (MUR) in the framework of the PON 2014-2021 “Research and Innovation" resources – Innovation Action - DM MUR 1062/2021 - Title of the Research: “Modelli semantici multimodali per l’industria 4.0 e le digital humanities.”, of PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - “FAIR - Future Artificial Intelligence Research" - Spoke 1 “Human-centered AI", funded by the European Commission under the NextGeneration EU programme and of the CrossLab and FoReLab projects (Departments of Excellence).

References

[1] A. Bondielli, F. Marcelloni, A survey on fake news and rumour detection techniques, Information Sciences 497 (2019) 38–55.

[2] P. Patwa, S. Sharma, S. Pykl, V. Guptha, G. Kumari, M. S. Akhtar, A. Ekbal, A. Das, T. Chakraborty, Fighting an infodemic: Covid-19 fake news dataset, in: Combating Online Hostile Posts in Regional Languages during Emergency Situation, Springer International Publishing, 2021, pp. 21–29.

[3] S. Lewandowsky, U. K. Ecker, J. Cook, Beyond misinformation: Understanding and coping with the “post-truth” era, Journal of Applied Research in Memory and Cognition 6 (2017) 353–369. URL: https://www.sciencedirect.com/science/article/pii/S2211368117300700. doi:https://doi.org/10.1016/j.jarmac.2017.07.008.

[4] F. Alam, S. Cresci, T. Chakraborty, F. Silvestri, D. Dimitrov, G. D. S. Martino, S. Shaar, H. Firooz, P. Nakov, A survey on multimodal disinformation detection, in: Proceedings of the 29th International Conference on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 6625–6643. URL: https://aclanthology.org/2022.coling-1.576.

[5] L. Bozarth, C. Budak, Toward a better performance evaluation framework for fake news classification, Proceedings of the International AAAI Conference on Web and Social Media 14 (2020) 60–71. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/7279.

[6] L. C. Passaro, A. Bondielli, A. Lenci, F. Marcelloni, Unipi-nle at checkthat! 2020: Approaching fact checking from a sentence similarity perspective through the lens of transformers, in: Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, 2020.

[7] G. Da San Martino, A. Barrón-Cedeño, H. Wachsmuth, R. Petrov, P. Nakov, SemEval-2020 task 11: Detection of propaganda techniques in news articles, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, ICCL, Barcelona (online), 2020, pp. 1377–1414. URL: https://aclanthology.org/2020.semeval-1.186. doi:10.18653/v1/2020.semeval-1.186.

[8] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, D. Testuggine, The hateful memes challenge: Detecting hate speech in multimodal memes, in: Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 2611–2624.

[9] D. Dimitrov, B. Bin Ali, S. Shaar, F. Alam, F. Silvestri, H. Firooz, P. Nakov, G. Da San Martino, SemEval-2021 task 6: Detection of persuasion techniques in texts and images, in: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), ACL, Online, 2021, pp. 70–98. URL: https://aclanthology.org/2021.semeval-1.7. doi:10.18653/v1/2021.semeval-1.7.

[10] F. Alam, A. Barrón-Cedeño, G. S. Cheema, S. Hakimov, M. Hasanain, C. Li, R. Míguez, H. Mubarak, G. K. Shahi, W. Zaghouani, P. Nakov, Overview of the CLEF-2023 CheckThat! lab task 1 on checkworthiness in multimodal and multigenre content, in: Working Notes of CLEF 2023—Conference and Labs of the Evaluation Forum, CLEF ’2023, Thessaloniki, Greece, 2023.

[11] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[12] A. Cossu, T. Tuytelaars, A. Carta, L. Passaro,

V. Lomonaco, D. Bacciu, Continual pre-training mitigates forgetting in language and vision, arXiv preprint arXiv:2205.09357 (2022).
[13] L. C. Passaro, A. Bondielli, P. Dell'Oglio, A. Lenci, F. Marcelloni, In-context annotation of topic-oriented datasets of fake news: A case study on the notre-dame fire event, Information Sciences (2022).
[14] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
[15] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[16] L. D'Amico, D. Napolitano, L. Vaiani, L. Cagliero, Polito at multi-fake-detective: Improving fnd-clip for multimodal italian fake news detection, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.
[17] Y. Zhou, Q. Ying, Z. Qian, S. Li, X. Zhang, Multimodal fake news detection via clip-guided learning, arXiv preprint arXiv:2205.14304 (2022).
[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.
[19] G. Puccetti, A. Esuli, Aimh at multi-fake-detective: System report, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.
[20] C. D. Hromei, D. Croce, V. Basile, R. Basili, Extremita at evalita2023: Multi-task sustainable scaling to large language models at its extreme, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.
[21] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, The Journal of Machine Learning Research 21 (2020) 5485–5551.
[22] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).
[23] S. Sarkar, N. Tudu, D. Das, Hijli-ju-clef at multi-fake-detective: Multimodal fake news detection using deep learning approach, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.
[24] Z. Gan, L. Li, C. Li, L. Wang, Z. Liu, J. Gao, Vision-language pre-training:: Basics, recent advances, and future trends, Foundations and Trends® in Computer Graphics and Vision 14 (2022) 163–352.
[25] V. Basile, Is EVALITA done? on the impact of prompting on the italian NLP evaluation campaign, in: Proceedings of the Sixth Workshop on Natural Language for Artificial Intelligence (NL4AI 2022), Udine, November 30th, 2022, volume 3287 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 127–140.
[26] L. C. Passaro, A. Lenci, Evaluating context selection strategies to build emotive vector space models, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), European Language Resources Association (ELRA), Portorož, Slovenia, 2016, pp. 2185–2191.

HIJLI-JU-CLEF at MULTI-Fake-DetectiVE: Multimodal Fake News Detection Using Deep Learning Approach

Sandip Sarkar[1,*], Nripen Tudu[2] and Dipankar Das[3]

[1]Department of Computer Science and Application, Hijli College, Kharagpur
[2]Department of Computer Science and Engineering, Jadavpur University, Kolkata
[3]Department of Computer Science and Engineering, Jadavpur University, Kolkata

Abstract

This report presents the progress made in developing a system for participation in a shared task called MULTI-Fake-DetectiVE Task 1, which aims to detect and verify fake news in a multimodal environment comprising both textual and visual elements. The task primarily revolves around automatically identifying fake news by analyzing the combined use of text and images. Our system is structured into three distinct modules. The initial module is responsible for extracting textual information from images. The second module serves as a translation component, enabling the analysis of non-English text by converting it into English. Lastly, the classification module utilizes the outputs from the previous modules to predict the appropriate classes, allowing for accurate differentiation between various types of content. Our objective is to create a system that can predict these labels. To achieve this, we extract information from both the image and the text, specifically focusing on English language text and translating any textual data into English. Subsequently, both sets of data are utilized to train a classification-based model, which aims to predict the aforementioned labels. We were able to obtain a Weighted Average F1-Score of 0.393 by implementing a Multi-head attention mechanism.

Keywords

Fake News, Fake news detection, Multi-modality, Vision-Language models, Large Language Models

1. Introduction

Over the past few years, there has been a substantial increase in Internet and social media usage. Unfortunately, this growth has been accompanied by a significant rise in the dissemination of fake news and misinformation. As a result, the ability to share information has become more accessible, no longer limited to prominent news organizations. Visual content, such as images, holds greater prominence on social media platforms due to its intuitive nature. The intentional dissemination of fake news aims to harm the reputation of individuals or organizations. It can serve as a propagandistic tool targeting political parties or specific communities. Fake news proliferates effortlessly across various platforms such as social media, online news platforms, blogs, messaging applications, and group conversations. To enhance its credibility and facilitate its dissemination on social media and other online channels, manipulated images and videos are frequently employed within fake news content. Certain websites and blogs employ deceitful designs and deceptive domain names to mimic legitimate news sources. Within closed groups and messaging apps, fake news can rapidly circulate among like-minded individuals through forwarded messages and posts. These techniques facilitate the swift dissemination of misinformation, further perpetuating false narratives.

The combination of natural language processing (NLP) and computer vision is mutually beneficial when it comes to detecting fake news and analyzing textual and visual content. NLP focuses on examining the language used in news articles and social media posts to detect patterns and inconsistencies. Identify and highlight untrue assertions, evaluate the reliability of sources, and verify information using reputable sources. Computer vision analyzes images and videos, detecting manipulated images by examining content, metadata, and contextual information. Detect indications of manipulation, deepfakes, or deceptive depictions. Combined, these methods provide a comprehensive approach to counteracting fake news.

Our paper exhibits a clear and logical organization, ensuring that the content is presented in a cohesive manner. In Section 2, we give a summary of related work means what other researchers have found in this area. Then, in Section 3, we provide a detailed explanation of MULTI-Fake-DetectiVE Task 1, using ideas and methods from previous research. We also look at the dataset for MULTI-Fake-DetectiVE Task 1 and share some interesting numbers and facts in Section 4. In the next part, which is Section 5, we go into detail about our model. Then, in Section 6, we describe our results. Finally, we wrap up our paper in Section 7 with our conclusions.

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
*Corresponding author.
sandipsarkar.ju@gmail.com (S. Sarkar);
nripentudu010@gmail.com (N. Tudu);
Dipankar.dipnil2005@gmail.com (D. Das)
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper33.pdf

2. Related work

Social media serves as a double-edged sword for news consumption. While it offers easy access to information at a low cost, it also enables the spread of fake news with intentionally false information [1]. Because there has been an increase in the use of text combined with images on social media, numerous studies have been conducted to examine how the combination of visual content and text can be used to anticipate the spread of false information.

In this research, the focus is on examining how the fusion of textual and visual content can be used to forecast deceptive information and false news. The study presents a multi-modal method that examines both text and images to detect patterns and discrepancies that suggest the presence of fake news [2].

Nguyen and Kyumin gathered and examined a group of online users known as guardians, who play a role in rectifying misinformation and fake news within online discussions by referencing fact-checking URLs. They introduced an innovative model for recommending fact-checking URLs, aiming to motivate guardians to actively participate in fact-checking endeavors [3].

The suggested framework employs the explicit convolution neural network model for image processing and the sentence transformer for text analysis. The features extracted from both visual and textual inputs are then fed through dense layers and subsequently merged to predict the authenticity of images [4].

3. Task Description

The primary focus of "MULTI-Fake-Detective Task 1 - Multimodal Fake News Detection and Verification" is to automatically identify fake news within environments that involve both text and images [1]. This task is framed as a multi-category classification problem within a multimedia context [5, 6].

The problem is defined as follows: when provided with a content element, denoted as c = ⟨t, v⟩, comprising a textual component t and a visual component v (such as an image), the task is to classify it into one of the following labels: "Certainly Fake," "Probably Fake," "Probably Real," or "Certainly Real."

These labels represent the overall information conveyed by a piece of content rather than its individual components. For instance, even if fake news includes authentic photographs presented in a deceptive context, it can still be considered misleading or entirely false. The labels can be interpreted as follows:

- **Certainly Fake:** news that is most certain to be fake, whatever the context.
- **Probably Fake:** news that is still likely to be fake, but may include some real information or at the very least be somewhat credible.
- **Probably Real:** news that is very credible but still retains some degree of uncertainty about the provided information.
- **Certainly Real:** news that is most certain to be real and incontestable, whatever the context.

4. Dataset Description

The dataset used for this project consists of a wide range of social media posts and news articles that include both text and images. These posts and articles are closely connected to real-world events that are often targeted for spreading false information. Specifically, the dataset focuses on the Ukrainian-Russian conflict that began in February 2022.

The provided download script was used to obtain the actual data. Table 1 shows the description of the dataset. This script creates a TSV (Tab-Separated Values) file and a directory named "Media" in the current working directory. The "Media" directory contains the downloaded images. The TSV includes the following:

I. **ID**: A unique identifier for the data point.
II. **URL**: The web address (URL) of the data point.
III. **Date**: The date when the data point was created. Please note that newspaper articles may not have this information available.
IV. **Type**: Indicates whether the data point is a news article or a tweet.
V. **Text**: The complete text of the data point.
VI. **Media**: The names of image files associated with the data point, if any.
VII. **Label**: A numerical label that represents one of the four possible labels (i.e. Certainly Fake: 0, Probably Fake: 1, Probably Real: 2, Certainly Real: 3) specified in the task description.

The “Media” directory contains all the images specified in the Media column of the TSV file.

5. System Description

The system consists of three modules designed to analyze both textual and visual components of content in order to classify it as fake or real. The First module of the system is the Image to Text generation module, which employs a pre-trained model to generate textual descriptions for images. The subsequent module is the translation module, which handles the task of translating text between

[1] https://sites.google.com/unipi.it/multi-fake-detective/home?authuser=0

Table 1
Data Points

ID	URL	Type	Label
1498022438398877704	https://twitter.com/manuela_carloni/status/1498022438398877704	tweet	1

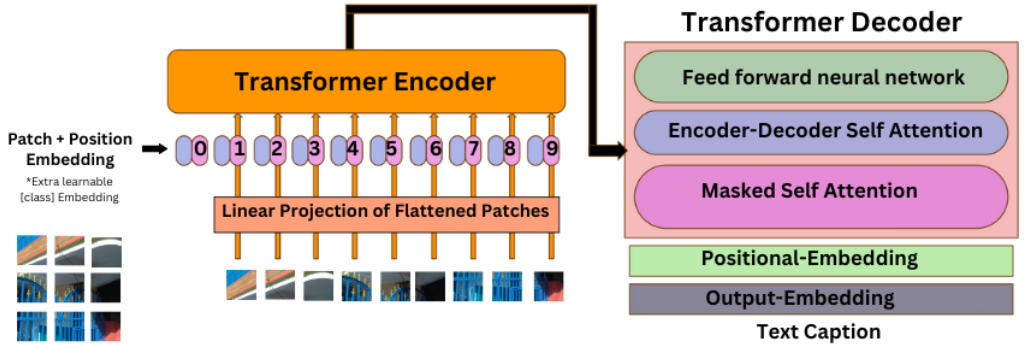


Figure 1: Image to Text Generation Model

different languages. The third module is dedicated to text classification and utilizes a combination of a BiLSTM (Bidirectional Long Short-Term Memory) model and a Multi-head attention model to classify text into different categories.

5.1. Image to Text generation Module

Image captioning is the process of generating a caption i.e., a description from the input image. It requires both Natural language processing as well as computer vision to generate the caption. Image captioning typically involves a deep learning approach, where a neural network model is trained on a large dataset of paired images and their corresponding captions.

This model learns to extract visual features from the images and then generates captions based on these visual features. The architecture of the Image to Text Generation Model is illustrated in Figure 1.

Image captioning is an example, in which the encoder model is used to encode the image, after which an autoregressive language model i.e., the decoder model generates the caption. The model we use to generate text from images is developed and trained by Ankur [2]. It is a Vision Encoder-Decoder Model, which is a type of neural network model used for tasks that involve processing both visual and textual information. It combines an encoder network, which processes the visual input, with a decoder network, which generates textual output. [3]

The encoder part of the model is responsible for extracting features from the visual input, such as an image. It typically consists of convolutional neural network (CNN) layers that can capture spatial and visual information from the input image. The encoder encodes the image into a fixed-length vector representation, often referred to as "image embedding" or "visual features."

The decoder part of the model takes the image embedding as input and generates textual output, such as captions or descriptions, related to the visual content. The decoder is typically implemented using recurrent neural networks (RNNs), such as long short-term memory (LSTM) or gated recurrent units (GRUs), which can process sequential data and generate coherent textual output.

The model was initialized with an image-to-text model with a pre-trained Transformer-based vision which involves dividing the input image into patches and treating them as tokens. These patches undergo self-attention and feed-forward networks within the Transformer encoder. Self-attention captures relationships between patches, while the feed-forward networks refine the representations. Positional embeddings capture spatial information, and a classification head predicts the final output. and a pre-trained language model as the decoder GPT2 (Generative Pre-trained Transformer 2) model operates through a mechanism called the Transformer architecture. It consists of multiple layers of self-attention and feed-forward neural networks. The model takes a sequence of tokens as input and processes them iteratively, attending to relevant tokens and generating context-aware representations. The self-attention mechanism captures relationships between different tokens in the sequence, while the feed-forward networks apply non-linear transforma-

[2]https://huggingface.co/nlpconnect/vit-gpt2-image-captioning
[3]https://ankur3107.github.io/assets/images/vision-encoder-decoder.png

tions.

The Vision Encoder-Decoder Model was trained using a dataset Common Objects in Context (COCO), a collection of more than 120 thousand images with descriptions [7]. The model is optimized to minimize the discrepancy between the predicted captions and the ground truth captions in the training data.

5.2. Link module

The Link module establishes a connection between Module 1 and Module 2. Since the data given by the organizer is in Italian, we need to convert it into a universal language, such as English. Subsequently, in the next module, we employ various classification methods.

5.3. Text Classification Module

The text classification module is a vital component of the multi-modal fake news detection system as it plays a pivotal role in analyzing the textual content and accurately categorizing it into one of the predefined labels: "Certainly Fake," "Probably Fake," "Probably Real," or "Certainly Real".

To perform text classification, we have employed two distinct models: 1) Bi-LSTM and 2) Multi-head attention model. Further information about these models can be found in the subsequent sections.

5.3.1. BiLSTM

BiLSTM, which stands for Bidirectional Long Short-Term Memory, is a recurrent neural network (RNN) model that takes into account information from both previous and future contexts when analyzing sequential data. Unlike traditional LSTM, which processes sequences in only one direction, BiLSTM processes the sequence in both forward and backward directions concurrently. It finds applications across various domains in the field of NLP, such as text simplification, machine translation, text similarity, and numerous other areas [8].

In the BiLSTM model, the input sequence is passed through two distinct LSTM layers: one layer handles the sequence in a forward direction, while the other layer handles it in a backward direction. This enables the model to grasp relationships and contextual information from both preceding and succeeding elements within the sequence.

Through the amalgamation of representations from both directions, BiLSTM adeptly captures a more holistic comprehension of the sequential data. This architecture is widely employed in various tasks, including natural language processing, speech recognition, and sentiment analysis, where incorporating context from both preceding and succeeding elements is crucial for precise predictions or classifications.

5.3.2. Multi-head Attention Mechanism

The Multi-head Attention Mechanism model is a component used in transformer-based neural networks that allows the model to attend to different parts of the input simultaneously and capture diverse relationships. It enhances the model's ability to process and extract information from the input sequence effectively.

In this mechanism, the input sequence is transformed into multiple query, key, and value representations through linear projections. These projections are then used to compute attention scores between different positions in the sequence. The attention scores determine the importance or relevance of each position with respect to others.

Multiple attention heads are employed in parallel, each attending to a different set of positions in the sequence. This enables the model to capture various patterns and dependencies at different levels of granularity. The output of each attention head is combined to form the final output, which incorporates information from multiple perspectives.

By utilizing the multi-head attention mechanism, the model can capture both local and global dependencies in the input sequence. It enhances the model's ability to model long-range dependencies, improve performance on complex tasks such as machine translation, text generation, and image captioning, and allows for efficient parallel computation during training and inference.

6. Results

Table 2 presents the rankings and performance of different teams in MULTI-Fake-DetectiVE Task 1, based on their weighted average F1 scores. The higher the F1 score, the better the performance of the team in that task. The "Weighted Avg. F1-Score" column displays the corresponding F1-score achieved by each team. The F1-score is a measure of a model's accuracy, combining precision and recall, and the weighted average takes into account the relative importance of each class in the evaluation.

7. Conclusion

In conclusion, the field of multi-modal fake news detection has witnessed significant research efforts aimed at detecting and verifying fake news in environments that involve both textual and visual elements. The studies mentioned above highlight the importance of considering multiple modalities, such as text and images, to enhance

Table 2
Result of MULTI-Fake-Detective Task 1 - Multimodal Fake News Detection

Rank	Team-Run	Weighted Avg. F1-Score
1	**Polito-P1**	**0.512**
2	extremITA-camoscio_lora	0.507
3	AIMH-MYPRIMARYRUN	0.488
4	Baseline-SVM_TEXT	0.479
5	Baseline-SVM_MULTI	0.463
6	Baseline-MLP_TEXT	0.448
7	Baseline-MLP_IMAGE	0.402
8	**HIJLI-JU-CLEF-Multi**	**0.393**
9	Baseline-SVM_IMAGE	0.386
10	Baseline-MLP_MULTI	0.374
11	**HIJLI-JU-CLEF-Bi-LSTM**	**0.314**

the accuracy and effectiveness of fake news detection systems.

One key aspect addressed by these studies is the combination of textual and visual content analysis to identify patterns, inconsistencies, and misleading information indicative of fake news. By analyzing the language used in news articles, social media posts, and the visual content accompanying them, researchers aim to detect and characterize fake news more comprehensively.

In conclusion, our system leveraged the combination of text and image analysis, employing state-of-the-art techniques in NLP and computer vision to detect and verify fake news in a multimodal environment. While we acknowledge the challenges associated with this task, our system demonstrates promising results and serves as a foundation for further advancements in combating fake news in multimedia contexts.

References

[1] K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, Fake news detection on social media: A data mining perspective, ACM SIGKDD Explorations Newsletter 19 (2017) 22–36.

[2] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, Y. Choi, Truth of varying shades: Analyzing language in fake news and political fact-checking, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 2931–2937. URL: https://aclanthology.org/D17-1317. doi:10.18653/v1/D17-1317.

[3] N. Vo, K. Lee, The rise of guardians: Fact-checking url recommendation to combat fake news, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 275–284. URL: https://doi.org/10.1145/3209978.3210037. doi:10.1145/3209978.3210037.

[4] B. Singh, D. K. Sharma, Predicting image credibility in fake news over social media using multi-modal approach, Neural Computing & Applications 34 (2022) 21503–21517. URL: https://doi.org/10.1007/s00521-021-06086-4. doi:10.1007/s00521-021-06086-4.

[5] A. Bondielli, P. Dell'Oglio, A. Lenci, F. Marcelloni, L. C. Passaro, M. Sabbatini, Multi-fake-detective at evalita 2023: Overview of the multimodal fake news detection and verification task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[6] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[7] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft COCO: common objects in context, CoRR abs/1405.0312 (2014). URL: http://arxiv.org/abs/1405.0312. arXiv:1405.0312.

[8] S. Sarkar, D. Das, P. Pakray, D. Pinto, A hybrid sequential model for text simplification, in: N. Priyadarshi, S. Padmanaban, R. K. Ghadai, A. R. Panda, R. Patel (Eds.), Advances in Power Systems and Energy Management, Springer Nature Singapore, Singapore, 2021, pp. 33–42.

AIMH at MULTI-Fake-DetectIVE: System Report

Giovanni Puccetti[1], Andrea Esuli[1]

[1]ISTI • Area della Ricerca CNR, via G. Moruzzi 1, 56124 Pisa, Italy

Abstract

This report describes our contribution to the EVALITA 2023 shared task MULTI-Fake-DetectIVE which involves the classification of news including textual and visual components. To experiment on this task we focus on textual data augmentation, extending the Italian text and the Images available in the training set using machine translation models and image captioning ones. To train using different set of input features, we use different transformer encoders for each variant of text (Italian, English) and modality (Image). For Task 1, among the models we test, we find that using the Italian text together with its translation improves the model performance while the captions don´t provide any improvement. We test the same architecture also on Task 2 although in this case we achieve less satisfactory results.

Keywords

MULTI-Fake-DetectIVE, Fake News, Multimodality

1. Introduction

Misinformation, intentional or not, is an ubiquitous phenomenon in social media. Whether due to malicious intent or scarce reviews, the number of outlets producing incorrect information is growing over time [1]. While the only true mean to protect one self from misinformation is careful review of trustworthy sources, the development of sound quantitative approaches for fake news detection is a worthy endeavour.

In this context there are works providing benchmark datasets for the very task of fake news detection in Twitter [2], however this is generally tackled in a unimodal setting where textual information is the only one examined. In this context, the MULTI-Fake-DetectIVE task [3], part of the EVALITA 2023 campaign [4] proposes to add multimodality, by challenging participants to classify fake news using both textual and visual features.

The task consists in classifying tweets reporting news about the war in Ukraine with both textual and visual content according to whether the reported news is true or fake. The task is subdivided into two subtasks:

- the first subtask is about detecting fake news by assigning a label among *Certainly False, Probably False, Probably True, Certainly True*;
- the second subtask is focused on detecting the agreement between text and image by assigning a label among *Misleading, Non Misleading, Unrelated*, which respectively indicate if the content of text and image support different interpretations, the same interpretation or are unrelated.

To perform the task we focus on exploring the effectiveness of augmenting the dataset by adding variants of the input extrapolated from both the existing text in Italian as well as the images leveraging the knowledge available in pre-trained models.

The idea of exploiting knowledge implicitly encoded in large pretrained models is used in several contexts with different goals, ranging from Neural Databases [5] to synthetic text detection [6].

The rest of the report is structured as follows: section 2 reports relevant literature, section 3 covers details of the dataset we found while preparing the models, section 4 is the System Description, section 5 outlines the results we obtained, and finally in section 6 we draw the conclusions of this work.

2. State of the Art

Recently, multimodal classification is tackled with visual language models such as OSCAR [7], VinVL [8] or with separate text and image encoding networks [9]. Built upon the idea of creating a shared representation space between text and images, developed in CLIP [10], several image captioning models have also been developed such as CoCa [11], we also try experimenting with these architecture for data augmentation. We could also use Multimodal Large Language Models for this same goal, i.e. augmenting data, some of the best performing ones are BLIP-2 [12] and Llava [13] these are too computationally costly and we avoid using them. Instead, to perform data augmentation across languages we employ Italian to English Neural Machine Translation models [14, 15].

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT

giovanni.puccetti@isti.cnr.it (G. Puccetti);
andrea.esuli@isti.cnr.it (A. Esuli)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper34.pdf

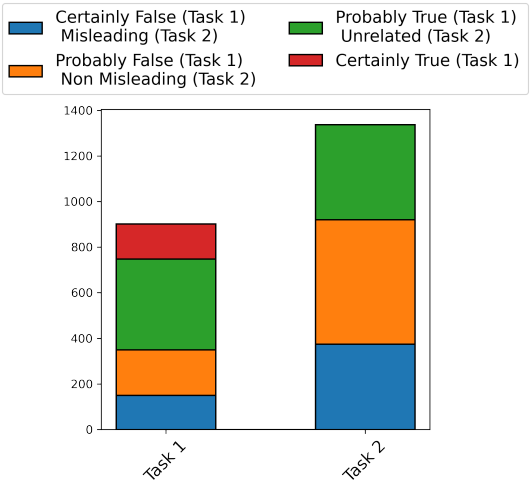


Figure 1: The labels distribution in the training dataset for Task 1 and 2.

3. Data

We perform an analysis of the dataset meant to understand if there are task specific preprocessing we have to apply to the data.

Figure 1 shows the distribution of labels in both tasks, we notice that both have heavily unbalanced distribution. The dataset of Task 1 Figure 1 shows how (likely) True news are the majority of samples, indeed, while ubiquitous in our everyday experience on the web, (likely) Fake news are still a minority of the total information shared.

Accordingly, for Task 2 Figure 1 shows that instances where Image and Text are heavily non aligned are also a minority.

While inspecting Task 1 training dataset, we observe non negligible data duplication, more specifically, there is 13.6% duplicated training samples, which we remove. On the contrary the dataset for Task 2 does not show any repetition.

4. Description of the System

In this Section we describe the methodology we developed to tackle the MULTI-Fake-DetectIVE task. We report the choices made and the steps that led us to them. In particular, we focus on data augmentation, for which we mainly adopt two systems working either on text or on images. Our architecture follows the one proposed by Gallo et al. [9].

We focus on data augmentation because the dataset is composed of Italian texts and since there aren't many models pre-trained specifically on this language we explore how well translating to English works. From here on, by *sample* we refer to a set of texts and images composing a single piece of news. Similarly, by *features* we indicate both texts or images.

To explore several data augmentation possibilities we build a unique pipeline that allows to add multiple pretrained models to process different input features schemes, based on different sets of texts and images.

Figure 2 outlines our architecture, using the same notation as in the figure, for each input sequence/image ($Feature_i$) in a sample, we use a pretrained model to embed it (N_i), then we add a linear layer ($Linear_i$) that maps all embeddings to the same dimension, finally we sum all such embeddings (entry-wise) to create a shared hidden state ($Hidden\ State$) and pass this vector through a linear layer ($Classification\ Head$) that maps it to a vector with length equal to the number of classes, 4 for task 1 and 3 for task 2. During training we optimize all parameters, including those of the pretrained models N_i.

4.1. Data Augmentation

The architecture we use allows us to seamlessly use as input any number of texts and images for each sample, in particular by adding extra features. We add features in two ways:

- We translate the textual documents to English using an open-source machine translation model [14, 15], in particular an Italian to English model[1];
- We caption the images using an image captioning model CoCa [11] fine tuned on the MSCOCO [16], we use an open source version [2].

Adding these extra inputs gives us the possibility to compose samples with different sets of features among *Italian Text, English Text, English Caption, Image.* We evaluate three sets of features:

- *English Text, Image*;
- *English Text, Italian Text, Image*;
- *English Text, English Caption, Image.*

4.2. Small Scale Ablation Study

All the models we test share the same high level architecture as shown in Figure 2, as mentioned above we use different pretrained transformer encoders to embed different modalities, sum all the embeddings entry wise after mapping them to the same dimension through a linear layer and finally with another linear layer we map to a vector with length the number of labels, finally we compute the usual Cross Entropy Loss for classification.

[1] www.huggingface.co/Helsinki-NLP/
[2] https://laion.ai/blog/coca/

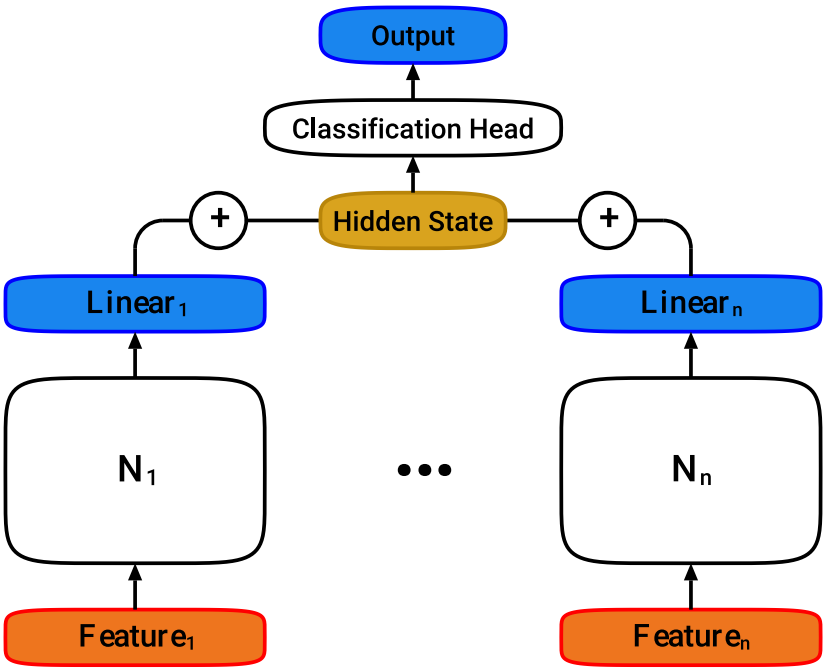


Figure 2: Neural Network architecture adaptable to several inputs schemes, N_i are pretrained transformers chosen specifically for each input augmented with a linear layer to map all output to the same dimension (512).

While summing the encoding of separate features, we multiply each of them by a coefficient, let us call it α_i (e.g. $\alpha_{eng-text}$ is the coefficient multiplying the embedding for the English translation), that modulates the relative importance of each feature. Similarly each feature has its own pre-trained encoder, we use the following ones:

- VIT [17] and in particular the vit-large-patch32-384 version[3] to encode images;
- RoBERTa large [18] to encode text in English, either the translated texts or the generated captions;
- a version of BERT-base pretrained on Italian[4] to encode all Italian text we use.

We perform all our validation test by splitting the training dataset in 80% training and 20% validation. The main architecture choices we make are, the shared size to which we map the embeddings output by each encoder, the α_i that multiply each of the embeddings before summing them, the classification head shape and the pre-trained models we use. Let us list how we chose each of them:

- For the vector size, we experiment with 512 and 1024, seeing that performance does not change depending on these two setting we use the smaller value, 512 in all our experiments.
- Concerning the α_i of each modality, we notice that $\alpha_{eng-text}$ is the most relevant one and after some tests, we choose the parameters as follows, $\alpha_{eng-text} = 1.0$ and all others equal to 0.1.
- The final ablation we have performed concerns the classification head, which eventually we choose to be a single linear layer with input size 512 and output size the number of labels, 4 for task 1 and 3 for task 2.
- Initially, we tried a different version with two Linear layers with *tanh* activation function in between and the hidden size of 2048, but this leads

[3] https://huggingface.co/google/vit-large-patch32-384
[4] https://github.com/dbmdz/berts

	Certainly Fake	Probably Fake	Probably Real	Certainly Real	weighted avg
support	16	52	106	21	195
precision	26.7	38.2	60.6	21.4	47.6
recall	25.0	25.0	75.5	14.3	51.3
f1-score	25.8	30.2	67.2	17.1	**48.6**

Table 1
Per class metrics using *English Text, Image* as features, our official task 1 submission (In **bold** the score used for the task).

	Misleading	Unrelated	Not Misleading	weighted avg
support	45	75	99	219
precision	28.3	46.8	47.6	43.3
recall	37.8	48.0	39.4	42.0
f1-score	32.4	47.4	43.1	**42.4**

Table 2
Per class metrics using *English Text, Image* as features, out official Task 2 submission (In **bold** the score used for the task).

to lower performance (although comparable) in all our experiments.

- Similarly, while choosing architecture we experimented with smaller versions of each transformer encoder, namely: (a) VIT with *patch16-224* instead of *patch32-384*; (b) roberta-base instead of roberta-large; (c) bert-base pretrained on English instead of Italian. However, while faster to train, switching any pretrained model to its smaller version reduced performance and therefore we opt for the larger ones when performing the grid search to choose our best model.

4.3. Hyper Parameter Selection

For Task 1 we perform a grid search using as features: *English Text, Image*.

We sweep over the following hyper parameters:

- learning rate: 1e-5, 2e-5, 3e-5, 5e-5;
- Max epochs: 3, 4, 5, 10, 20;
- Warmup steps: 0, 100;
- Batch size: 4, 8 (other values would not fit into our machine).

The best performance on our validation set is obtained with warmup 0, batch-size 8, epochs 4 and learning rate 1e-5 and therefore we use this set of hyper parameters when training with all groups of features[5].

Due to limitations in GPU memory, we clip all sequences to 256 length. We also tested a length of 400 using English Text and Images only, however this did not seem to affect performance[6].

Comparing the results obtained on our validation set when using different groups of features we eventually choose to only use the translated text together with the Images, as adding Italian didn't appear to provide significant improvements.

We tackle Task 2 keeping everything as we did in Task 1 switching training set.

5. Results

5.1. Task 1

Table 1 shows the performance of our approach on the first task. In bold we report the metric that has been used to evaluate our model, it reports how the class balance in the training set is reflected into per-class performance into the official test set (measured with the official evaluation script). Indeed the *Certainly Real* class is the most numerous in this case too as well as the one where our model is best performing. It is interesting to notice how the model performs better on the *Certainly False* class compared to the *Certainly Real* one despite the second being more populated, we speculate this is due to the similarity with the *Probably Real* class.

Although we chose a different method to submit to Task 1, we show that including the Italian text results into promising results on the official test set. Table 3 shows how this approach performs on the official test set and indeed it would improve over our submission.

Unlike adding the Italian text, using the captions does not results in performance improvements. Table 4 shows the performance obtained while adding the Captions for the images, generated by CoCa [11] and processed with a different roberta-large model.

5.2. Task 2

For task 2, we chose to keep all parameters as in Task 1.

[5]Performing a separate grid search for each feature group was not feasible.

[6]Other configurations using more features do not fit in memory for length of 400. All experiments are performed on NVIDIA GTX Quadro with 24GB of VRM.

	Cert. Fake	Prob. Fake	Prob. Real	Cert. Real	weight. avg
support	16	52	106	21	195.0
precision	50.0	41.2	62.9	20.0	51.4
recall	43.8	26.9	78.3	14.3	54.9
f1-score	46.7	32.6	69.7	16.7	**52.2**

Table 3
Per class metrics for Task 1 with features *English Text, Italian Text, Image* (In **bold** the score used for the task).

	Cert. Fake	Prob. Fake	Prob. Real	Cert. Real	weigh. avg
support	16	52	106	21	195
precision	12.5	46.2	59.6	20.0	47.9
recall	6.2	23.1	82.1	14.3	52.8
f1-score	8.3	30.8	69.0	16.7	**48.2**

Table 4
Per class metrics for Task 1 with features *English Text, English Captions, Image* (In **bold** the score used for the task).

Table 2 shows the scores we obtain on the official Test set for Task 2, it appears that the model we use can´t recognize the *Misleading* class for which f1-score is 32.4% while it does manage to achieve higher values on the remaining classes. We conclude that the relation between the Text and Image embeddings are not well captured by this model.

6. Conclusions

We have tackled the MULTI-Fake-DetectIVE task trying and improve performance with textual data augmentation techniques.

We show that our approach does provide some improvements and this is relevant as text-based data augmentation is a novel way to exploit the knowledge present within large pretrained models, made recently possible by pretrained models and has several application settings [19].

Moreover, in this report we show how using both Italian and English data at once, even though the English one is the translation of the Italian text, provides significant improvements in Task 1.

On the contrary, the lower performance of the model in Task 2 underlines how the relations between text and images are not well captured by our model and this offers the opportunity for further improvements.

A structural limitation of our approach is that, although we know that the dataset is composed of both tweets and articles and that the second document type is generally much longer than the tweets, we have not experimented with ways to use this longer context.

This too offers a promising future step, using longer context transformers when embedding text, while keeping our overall scheme of translating to English might give further improvements.

Indeed, given the scarcity of longer context transformers trained on Italian the English translation might be useful in this case as well.

Acknowledgments

This work is supported by the European Union under the scheme HORIZON-INFRA-2021-DEV-02-01 – Preparatory phase of new ESFRI research infrastructure projects, Grant Agreement n.101079043, "SoBigData RI PPP: SoBigData RI Preparatory Phase Project"

References

[1] X. Zhou, R. Zafarani, A survey of fake news: Fundamental theories, detection methods, and opportunities, ACM Comput. Surv. 53 (2020). URL: https://doi.org/10.1145/3395046. doi:10.1145/3395046.

[2] T. Fagni, F. Falchi, M. Gambini, A. Martella, M. Tesconi, Tweepfake: About detecting deepfake tweets, PLOS ONE 16 (2021) 1–16. URL: https://doi.org/10.1371/journal.pone.0251415. doi:10.1371/journal.pone.0251415.

[3] A. Bondielli, P. Dell'Oglio, A. Lenci, F. Marcelloni, L. C. Passaro, M. Sabbatini, Multi-fake-detective at evalita 2023: Overview of the multimodal fake news detection and verification task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[4] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[5] J. Thorne, M. Yazdani, M. Saeidi, F. Silvestri, S. Riedel, A. Halevy, From natural language processing to neural databases, Proc. VLDB Endow. 14 (2021) 1033–1039. URL: https://doi.org/10.14778/3447689.3447706. doi:10.14778/3447689.3447706.

[6] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, C. Finn, Detectgpt: Zero-shot machine-generated text detection using probability curvature, 2023. arXiv:2301.11305.

[7] X. Li, X. Yin, C. Li, X. Hu, P. Zhang, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, J. Gao, Oscar: Object-semantics aligned pre-training for vision-language tasks, ECCV 2020 (2020).

[8] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, J. Gao, Vinvl: Making visual representations matter in vision-language models, CVPR 2021 (2021).

[9] I. Gallo, A. Calefati, S. Nawaz, M. K. Janjua, Image and encoded text fusion for multi-modal classification, 2018 Digital Image Computing: Techniques and Applications (DICTA) (2018) 1–7.

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 8748–8763. URL: https://proceedings.mlr.press/v139/radford21a.html.

[11] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, Y. Wu, Coca: Contrastive captioners are image-text foundation models, 2022. arXiv:2205.01917.

[12] J. Li, D. Li, S. Savarese, S. C. H. Hoi, BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models, CoRR abs/2301.12597 (2023). URL: https://doi.org/10.48550/arXiv.2301.12597. doi:10.48550/arXiv.2301.12597. arXiv:2301.12597.

[13] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, J. Gao, Llava-med: Training a large language-and-vision assistant for biomedicine in one day, CoRR abs/2306.00890 (2023). URL: https://doi.org/10.48550/arXiv.2306.00890. doi:10.48550/arXiv.2306.00890. arXiv:2306.00890.

[14] J. Tiedemann, Parallel data, tools and interfaces in OPUS, in: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey, 2012.

[15] J. Tiedemann, S. Thottingal, OPUS-MT — Building open translation services for the World, in: Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT), Lisbon, Portugal, 2020.

[16] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Doll'a r, C. L. Zitnick, Microsoft COCO: common objects in context, CoRR abs/1405.0312 (2014). URL: http://arxiv.org/abs/1405.0312. arXiv:1405.0312.

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.

[18] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Ro{bert}a: A robustly optimized {bert} pretraining approach, 2020. URL: https://openreview.net/forum?id=SyxS0T4tvS.

[19] A. Mumuni, F. Mumuni, Data augmentation: A comprehensive survey of modern approaches, Array 16 (2022) 100258. URL: https://www.sciencedirect.com/science/article/pii/S2590005622000911. doi:https://doi.org/10.1016/j.array.2022.100258.

PoliTo at MULTI-Fake-DetectiVE: Improving FND-CLIP for Multimodal Italian Fake News Detection

Lorenzo D'Amico[1], Davide Napolitano[1], Lorenzo Vaiani[1] and Luca Cagliero[1]

[1]Politecnico di Torino, Turin, Italy

Abstract

The MULTI-Fake-DetectiVE challenge addresses the automatic detection of Italian fake news in a multimodal setting, where both textual and visual components contribute as potential sources of fake content. This paper describes the PoliTO approach to the tasks of fake news detection and analysis of the modality contributions. Our solution turns out to be the best performer on both tasks. It leverages the established FND-CLIP multimodal architecture and proposes ad hoc extensions including sentiment-based text encoding, image transformation in the frequency domain, and data augmentation via back-translation. Thanks to its effectiveness in combining visual and textual content, our solution contributes to fighting the spread of disinformation in the Italian news flow.

Keywords

Fake News Detection, Multimodal Learning

1. Introduction

In recent years, the proliferation of fake news and disinformation in online platforms has become a significant challenge, impacting public discourse, political landscapes, and social dynamics [1]. This phenomenon has been particularly evident during real-world events, where misinformation spreads rapidly, often leading to harmful consequences.

With the proliferation of multimodal data sources, detecting fake news content has become more and more challenging as fake content can be hidden in either visual and textual news elements. Studying the interplay between these modalities is crucial for understanding how misinformation is crafted and disseminated and how it can be effectively detected.

The MULTI-Fake-DetectiVE challenge [2] proposed at EVALITA 2023 [3] focuses on overcoming the limitations of existing approaches in coping with multimodal Italian news content. It addresses the automatic detection of Italian fake news in a multimodal setting, where both textual and visual components potentially contribute as sources of fake content. The challenge has the twofold aim of accurately discriminating between real and fake news content and investigating the influence of visual and textual components on each other's interpretation.

In this work, we present the PoliTO approach to both MULTI-Fake-DetectiVE tasks. Our solution outperforms all the baselines and competitors in both tasks, strengthening the state-of-the-art multimodal fake news detectors in a challenging and previously underexplored scenario. Furthermore, it releases new open-source resources to the community to fight disinformation in the Italian flow of news articles.

We propose FND-CLIP-IT, an extension of the state-of-the-art FND-CLIP multimodal architecture. We explore the integration of a sentiment-based text encoder, a data augmentation stage based on back-translation, an image transformation module based on Discrete Fourier Transform, as well as different approaches to combine and weigh the input embeddings.

The remainder of this paper is organized as follows. In Section 2 we review the literature on fake news detection, considering both text-only and multimodal approaches. Section 3 briefly describes the dataset, task, and metrics used in the challenge. In Section 4 we describe the methodology, primarily focusing on the proposed FND-CLIP extensions. Section 5 presents the experimental setup and the obtained results. Finally, Section 6 draws the conclusions and discusses the main limitations and future directions.

2. Related Work

NLP-based approaches. Early approaches focused on linguistic features, such as lexical and syntactic patterns, to distinguish between real and fake news. However, with the advancement of deep learning researchers have increasingly turned to more sophisticated methods such as recurrent neural networks [4], convolutional neural networks [5], and transformer models [6] to capture semantic and contextual information for improved

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7–8, Parma, IT

lor.damico@studenti.polito.it (L. D'Amico); davide.napolitano@polito.it (D. Napolitano); lorenzo.vaiani@polito.it (L. Vaiani); luca.cagliero@polito.it (L. Cagliero)

0000-0001-9077-4103 (D. Napolitano); 0000-0002-3605-1577 (L. Vaiani); 0000-0002-7185-5247 (L. Cagliero)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper35.pdf

detection accuracy. In this work, we mainly rely on state-of-the-art transformers pretrained on Italian textual data to effectively extract information from the news textual component.

Multimodal Approaches. Incorporating multimodal information such as text and images has shown to be promising to improve the accuracy of fake news detection systems [7]. Recently, the adoption of multimodal architectures and transformers has shown to be particularly effective in capturing the semantic relationships among different modalities for fake news detection, e.g., CB-Fake [8], CAFE [9] and TTEC [10].

FND-CLIP, proposed by [11], is among the most recently proposed multimodal architectures for fake news detection. It relies on the established CLIP model [12] to measure the cross-modal similarity and guide the mapping and fusion of the input features. The architecture develops along three main streams: a textual one, which extracts information using BERT and CLIP, a visual one which extracts features from the images using ResNet and CLIP, and a multimodal one which combines the features extracted using CLIP from both modalities. FND-CLIP suffers from the following limitations:

- The natural language encoder neglects the polarity of the input text, which is known to be relevant to fake news detection [13].
- Fake news examples are likely to be undersampled in real training data. Hence, the classification model may suffer from class imbalance effects.
- Multimodal fake news often contains tampered visual content. Tampered images are more likely to be detected in the frequency domain space. However, FND-CLIP does not consider any frequency-based image descriptor [14].

Our research endeavors to address the aforesaid limitations by proposing FND-CLIP-IT, i.e., an improved version FND-CLIP suited to multimodal Italian fake news detection.

3. Task and Dataset Description

3.1. Tasks Description

Task 1: Multimodal Fake News Detection. The problem can be formulated as a multi-class classification task, where the input content $c = \langle t, v \rangle$, consisting of a textual component t and a visual component v, can be classified as follows: *Certainly Fake* (CF), when the news is very likely to be fake, regardless of the context in which it is presented; *Probably Fake* (PF), when the news is likely to be fake but may contain some real information or exhibit a certain level of credibility; *Probably Real* (PR), when the news is highly credible but retains some degree of uncertainty regarding the information provided; *Certainly Real* (CR), when the news is most certain to be real and indisputable, regardless of the context. It is worth noticing that these labels pertain to the overall informational content and should not be assigned based solely on the individual components.

Task 2: Analysis of Cross-Modal Relations in Fake and Real News. The purpose is to examine the relationship between the textual and visual modalities within the context of fake and real news. The primary objective is to gain insights into how images and texts in fake and real news can potentially lead to misleading interpretations of the content, both within each modality and as a whole. The task can be formulated as a three-class classification problem. Given a multimodal piece of content c, the goal is to automatically assign one of the following categories to c: *Misleading* (M), when either the image or the text is misleading in terms of interpreting the information conveyed by the other modality or the content as a whole; *Not Misleading* (NM), when both image and text are related to each other, providing support to the overall information presented, and are not intended to mislead; *Unrelated* (U), when the image and the text are not related to each other.

Evaluation Metrics. For both tasks, the evaluation metrics are accuracy, average per-class precision and recall, and macro- and weighted- F1 score. Weighted-F1 score has been selected as the reference metric for ranking participants.

3.2. Dataset Description

Both task-specific datasets consist of a collection of Twitter posts and newspaper articles describing one or more real events. For Task 1 the training set contains 908 distinct labeled samples[1]. The labels in the training data are distributed as follows: CF 16.4%, PF 22.0%, PR 44.4%, CR 17.2%. Around 80.0% of the samples are tweets, whereas the remaining ones are news articles. The test set consists of 193 samples following roughly the same per-class and per-type distributions as in the training data. For Task 2, the training set contains 1309 distinct samples and the per-class distribution is M 26.9%, U 40.6%, NM 31.5%. 66.0% of the samples are tweets, whereas the remaining ones are news articles. The test set contains 219 samples. Compared to the training data, the per-type sample distribution is slightly more biased towards tweets (75.0%) and Non-Misleading content (45.2%).

[1] available at the time of writing, June 2023

4. Methodology

Here we present FND-CLIP-IT, an improved version of FND-CLIP suited to the MULTI-Fake-DetectiVE challenge. Our solution is rooted in the original FND-CLIP model [11] and a set of unimodal language and visual encoders described below.

Unimodal language baselines. We utilize the following models tailored to the Italian language: BERT-IT[2], GilBERTo[3], BART-IT[4] [15].

Since the input text can be longer than the maximum model size, we adopt a hierarchical approach: the text is divided into chunks of fixed length, where each one is fed to the transformer encoder, and then the final representation is obtained by averaging all the [CLS] tokens.

Unimodal visual baselines. We exploit two established models, i.e., ViT [16] and ResNet-152 [17]. Since more pictures can be associated with the same sample, at inference time we separately evaluate all the images and the final prediction is the average of all obtained output logits.

Multimodal baselines. To leverage visual and textual content at the same time, we rely on (i) the standard FND-CLIP [11] architecture, adapted to handle Italian text rather than English, (ii) CLIP [12], and (iii) a late fusion approach combining BERT-IT and ResNet-152.

4.1. FND-CLIP-IT

FND-CLIP-IT extends the state-of-the-art FND-CLIP architecture to address the current limitations of fake news detection approaches. By incorporating the proposed extensions, the overall efficacy and robustness of FND-CLIP-IT shows significant improvements compared to the baseline versions. A detailed description of the proposed extensions, hereafter denoted by *A*, *B*, *C*, *D*, and *E* for the sake of brevity, is given below.

A. **Sentiment-based textual representation**: To consider the polarity of the input text for fake news detection [13], we enrich the textual representation by adding a sentiment-based encoding to the existing text encoders. Specifically, we use the Italian-BERT model finetuned on a sentiment analysis task[5]. We also consider the following variants of sentiment-based textual representation:
   - $\mathbf{A}_1$. A concatenation of the sentiment-based embedding to the initial original representation, on top of which we apply the textual projection head.
   - $\mathbf{A}_2$. A separate stream of information with a dedicated projection head.

B. **DFT-based additional stream**: we convert the image from the spatial domain to the frequency domain by applying Discrete Fourier Transform (DFT). The purpose is to detect tampered images, which likely occur in multimodal fake news [14]. We encode both real and imaginary parts using a dedicated VGG19 [18]. The obtained representations are then concatenated to generate a parallel stream of information that will be then combined with the others before applying the final FND-CLIP classifier.

C. **Embedding concatenation**: instead of summing the embedding of each stream we concatenate them. Concatenation has already been proven to be an effective way of combining multimodal information [19]. The rationale behind it is that by keeping more fine-grained pieces of information the classification head, adapted to handle the new encoding, can capture the most discriminating source features in a more effective way.

D. **Class rebalancing through data augmentation**: since the dataset is quite imbalanced across the classes, we re-balance the data distribution by penalizing the most frequent class. In particular, we generate new samples of the minority classes by applying a textual augmentation based on back-translation [20], which already proved to be beneficial in both multimodal [21] and fake news detection [22] tasks. The auxiliary language adopted is English, and the translation models used are provided by Helsinki-NLP[6].

E. **Additional Squeeze and Excitation Layers**: Similar to FND-CLIP, we employ a squeeze-and-excitation operation [23] to weigh the input embedding streams. The purpose of a squeeze-and-excitation block is to adaptively recalibrate channel-wise feature responses by explicitly modeling interdependencies between channels. Unlike [11], where they weigh differently the textual and visual streams, we also adopt squeeze-and-excitation within each modality to weigh the relevance of each encoder. The key idea is to give more importance to discriminating modality-specific embeddings.

Beyond considering each FND-CLIP extension separately, we also build both models that combine the pro-

[2] https://huggingface.co/dbmdz/bert-base-italian-xxl-cased

[3] https://huggingface.co/idb-ita/gilberto-uncased-from-camembert

[4] https://huggingface.co/morenolq/bart-it

[5] https://huggingface.co/neuraly/bert-base-italian-cased-sentiment

[6] https://huggingface.co/Helsinki-NLP

Model	Accuracy	Precision	Recall	F1-macro	F1-weighted
BERT-IT	0.554	0.558	0.470	0.491	0.531
GilBERTo	0.522	0.510	0.462	0.473	0.514
BART-IT	0.495	0.447	0.428	0.429	0.472
ResNet-152	0.451	0.406	0.373	0.380	0.436
ViT	0.402	0.336	0.322	0.324	0.388
BERT-IT+ResNet-152	0.516	0.516	0.423	0.433	0.479
CLIP-IT	0.538	0.523	0.478	0.484	0.520
FND-CLIP-IT	0.560	0.551	0.492	0.503	0.537
FND-CLIP-IT$_{A_1}$	0.565	0.558	0.525	0.530	0.552
FND-CLIP-IT$_{A_2}$	0.560	0.541	0.521	0.526	0.550
FND-CLIP-IT$_{B}$	0.587	0.587	0.506	0.525	0.565
FND-CLIP-IT$_{C}$	0.576	0.578	0.521	0.539	0.568
FND-CLIP-IT$_{D}$	0.565	0.540	0.517	0.524	0.555
FND-CLIP-IT$_{E}$	0.576	0.563	0.558	0.560	0.574
FND-CLIP-IT$^{*}_{A_2,B}$	0.587	0.578	0.548	0.560	0.581
FND-CLIP-IT$^{*}_{A_1,C,D}$	<u>0.609</u>	0.595	<u>0.593</u>	<u>0.593</u>	<u>0.606</u>
FND-CLIP-IT$^{*}_{A_1,D,E}$	0.598	<u>0.603</u>	0.569	0.582	0.596
FND-CLIP-IT$_{A_1,B,C,D,E}$	0.576	0.568	0.529	0.542	0.572
ENSEMBLE	**0.658**	**0.661**	**0.621**	**0.637**	**0.653**

Table 1
Results obtained on the validation set from our baselines, FND-CLIP-IT variants and ensemble models. The best results are highlighted in bold, while the best results of a single model are underlined. Models marked with * are involved in the reported ensemble.

posed extensions *A-E* in different ways and ensemble methods that combine best-performing individual models. To this end, we use a weighted average of individual logits for each class.

5. Experimental Results

5.1. Setup

The models were fine-tuned for a maximum of 80 epochs, using a batch size of 16, a learning rate of 1e-3, an AdamW optimizer with a weight decay of 0.001 and a linear scheduler. All baseline models were trained using a cross-entropy-loss, while all FND-CLIP-IT variants were trained with both cross-entropy and focal losses.

5.2. Results

Table 1 presents the results of the baselines (upper part) and our proposed solutions (lower part), obtained on the Task 1 validation set. Significantly, the outcomes reveal an intriguing pattern wherein text-only models exhibit superior performance when compared to image-only models, underscoring the paramount importance of textual information within the context of the task. Notably, the multimodal CLIP baseline demonstrates results comparable to the text-only model. At the same time, FND-CLIP-IT architecture attains performance marginally better than the BERT-IT model. Furthermore, our diverse extensions of the FND-CLIP-IT framework, when applied individually, yield notable improvements over the original implementation. In addition, select combinations of these variants produce even more promising outcomes. Although both focal loss and cross-entropy were evaluated, we chose to report only the results obtained with focal loss, due to their overall superior performance compared to cross-entropy. It is worth noting, however, that the combination of all variants does not surpass the performance of specific combinations, indicating a potential susceptibility to overfitting. Furthermore, an intriguing observation emerges with the implementation of an ensemble model that leverages the best-performing combinations. This ensemble model outperforms the individual models, further accentuating the benefits of employing ensemble techniques to enhance overall performance.

5.3. Competition

We employed our ensemble method to evaluate the performance of our FND-CLIP-IT variants on the test samples. The test results are presented in Table 2. The upper part of Table 2 shows the outcomes obtained for Task 1. Although these results are worse than the performance achieved on our validation set, they surpass all other baselines and competitors.

Furthermore, we fine-tuned the same ensemble of models for Task 2 by replacing the classification head last layer. The bottom of Table 2 reports the achieved results for Task 2. Remarkably, our approach outperforms both the baseline and the competitors.

	Team Run	F1-weighted
Task 1	PoliTo-P1	**0.512**
	extremITA-camoscio_lora	0.507
	AIMH-MYPRIMARYRUN	0.488
	Baseline-SVM-TEXT	0.479
	HIJLI-JU-CLEF-Multi	0.393
Task 2	PoliTo-P1	**0.517**
	Baseline-MLP-TEXT	0.506
	AIMH-MYPRIMARYRUN	0.421

Table 2
Official MULTI-Fake-DetectiVE results. For the official baselines, we report only the result of the best-performing approach.

By leveraging our best ensemble method, we have demonstrated the robustness and versatility of our FND-CLIP-IT variants across both Task 1 and Task 2, surpassing existing approaches in terms of performance and effectiveness.

6. Conclusion and Future Directions

In this study, we introduced the FND-CLIP-IT architecture exploring several variants for fake news detection in a multimodal setting. Our findings demonstrate the effectiveness of these variants, with notable improvements observed over the original implementation. Furthermore, selected combinations of these variants and model ensembling yield even more promising results.

In the future, we plan to continue refining and optimizing the proposed variants to further enhance their performance. Additionally, investigating the interpretability of the FND-CLIP-IT architecture and understanding its decision-making process will be an interesting direction for future research.

Acknowledgments

This study was carried out within the FAIR - Future Artificial Intelligence Research and received funding from the European Union Next-GenerationEU (PNRR M4C2, INVESTIMENTO 1.3 D.D. 1555 11/10/2022, PE00000013). This study was carried out within the MICS (Made in Italy – Circular and Sustainable) Extended Partnership and received funding from the European Union Next-GenerationEU (PNRR M4C2, INVESTIMENTO 1.3 D.D. 1551.11-10-2022, PE00000004). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them. The research leading to these results has been partly funded by the SmartData@PoliTO center for Big Data and Machine Learning technologies. Computational resources were provided by HPC@POLITO, a project of Academic Computing within the Department of Control and Computer Engineering at the Politecnico di Torino[7].

References

[1] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, Science (2018).

[2] A. Bondielli, P. Dell'Oglio, A. Lenci, F. Marcelloni, L. C. Passaro, M. Sabbatini, Multi-fake-detective at evalita 2023: Overview of the multimodal fake news detection and verification task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[3] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[4] C. Iwendi, S. Mohan, S. khan, E. Ibeke, A. Ahmadian, T. Ciano, Covid-19 fake news sentiment analysis, Computers and Electrical Engineering (2022).

[5] K. Ma, C. Tang, W. Zhang, B. Cui, K. Ji, Z. Chen, A. Abraham, Dc-cnn: Dual-channel convolutional neural networks with attention-pooling for fake news detection, Applied Intelligence (2023).

[6] H. Jwa, D. Oh, K. Park, J. M. Kang, H. Lim, exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert), Applied Sciences 9 (2019).

[7] I. Segura-Bedmar, S. Alonso-Bartolome, Multimodal fake news detection, Information 13 (2022).

[8] B. Palani, S. Elango, V. Viswanathan K, Cb-fake: A multimodal deep learning framework for automatic fake news detection using capsule neural network and bert, Multimedia Tools and Applications (2022).

[9] Y. Chen, D. Li, P. Zhang, J. Sui, Q. Lv, L. Tun, L. Shang, Cross-modal ambiguity learning for multimodal fake news detection, in: Proc. of the ACM Web Conference 2022, 2022, pp. 2897–2905.

[10] J. Hua, X. Cui, X. Li, K. Tang, P. Zhu, Multimodal fake news detection through data augmentation-based contrastive learning, Applied Soft Computing 136 (2023) 110125.

[11] Y. Zhou, Q. Ying, Z. Qian, S. Li, X. Zhang, Multimodal fake news detection via clip-guided learning, arXiv preprint arXiv:2205.14304 (2022).

[7] https://www.hpc.polito.it/

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021.
[13] M. A. Alonso, D. Vilares, C. Gómez-Rodríguez, J. Vilares, Sentiment analysis for fake news detection, Electronics 10 (2021).
[14] J. Jing, H. Wu, J. Sun, X. Fang, H. Zhang, Multimodal fake news detection via progressive fusion networks, Information Processing & Management (2023).
[15] M. La Quatra, L. Cagliero, Bart-it: An efficient sequence-to-sequence model for italian text summarization, Future Internet 15 (2023).
[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: 9th International Conference on Learning Representations, ICLR 2021, 2021.
[17] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. of the IEEE conference on computer vision and pattern recognition, 2016.
[18] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Proc. of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, 2015.
[19] L. Vaiani, M. La Quatra, L. Cagliero, P. Garza, Leveraging multimodal content for podcast summarization, in: Proc. of the 37th ACM/SIGAPP Symposium on Applied Computing, 2022.
[20] S. Edunov, M. Ott, M. Auli, D. Grangier, Understanding back-translation at scale, in: Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing, ACL, Brussels, Belgium, 2018.
[21] L. Vaiani, L. Cagliero, P. Garza, PoliTo at SemEval-2023 task 1: Clip-based visual-word sense disambiguation based on back-translation, in: Proc. of the 17th International Workshop on Semantic Evaluation (SemEval-2023), ACL, Toronto, Canada, 2023.
[22] M. Amjad, G. Sidorov, A. Zhila, Data augmentation using machine translation for fake news detection in the urdu language, in: Proc. of the 12th language resources and evaluation conference, 2020, pp. 2537–2542.
[23] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proc. of the IEEE conference on computer vision and pattern recognition, 2018.

ACTI at EVALITA 2023: Automatic Conspiracy Theory Identification Task Overview

Giuseppe Russo[1], Niklas Stoehr[1] and Manoel Horta Ribeiro[2]

[1]ETH Zürich, Rämistrasse 101, 8092 Zürich, Switzerland
[2]EPFL, Rte Cantonale, 1015 Lausanne, Switzerland

Abstract

English. Automatic Conspiracy Theory Identification (ACTI) is a new shared task proposed for the first time at the EVALITA 2023 evaluation campaign. ACTI is based on a new, manually labeled dataset of comments scraped from conspiratorial Telegram channels and consists of two subtasks: (1) identifying conspiratorial content (conspiratorial content classification); and (2) classifying content into specific conspiracy theories (conspiratorial category classification). A total of 15 teams participated in the task with 81 submissions. In this task summary, we discuss the data and task, and outline the best-performing approaches that are largely based on large language models. We conclude with a brief discussion of the application of large language models to counter the spread of misinformation on online platforms.

Keywords

Conspiracy Theory, Content Moderation, Large Language Models, Computational Social Science

1. Introduction

From ancient tales of secret societies, [2] to speculation on whether the moon landing happened [3], belief in conspiracy theories has been prevalent throughout human history [4] and has inflicted harm upon individuals and groups falsely accused of wrongdoing [5]. For example, in the middle ages, the Blood Libel conspiracy theory falsely accused Jews of murdering Christian boys, fostering their persecution [6].

Fast-forward to the digital age, the Internet has emerged as the prominent medium through which individuals are exposed to conspiracy theories [7, 8]. Indeed, mainstream and fringe platforms have served as *de-facto* incubators of online conspiracies [9]. Notably, the impact of online conspiracy theories has been far-reaching, inciting real-world violence and influencing public health. The QAnon conspiracy, which gained momentum during the Trump administration, was pivotal in planning the 2021 invasion of the US Capitol [10, 11]. At the same time, the conspiracy theories associated with COVID-19 fueled anti-vaccination sentiments and skepticism towards public health measures [12, 13].

Mainstream platforms limit the diffusion of conspiratorial content through interventions that range from banning online communities [14, 15] to telling users that the information presented may be inaccurate [16]. While these interventions may help curb the proliferation of conspiracy theories in online spaces [17], they require a fundamental technology: ways to identify conspiratorial content accurately and at scale across various languages and cultural contexts [18].

In this context, we propose the *Automatic Conspiracy Theory Identification (ACTI)* task. Considering a dataset with over 25 thousand posts in Italian extracted from five Telegram channels, the ACTI consists of two subtasks: (i) a binary classification task where the goal is to determine if a given text piece is conspiratorial or not; and (ii) a multi-class classification task to recognize specific conspiracy theories.

2. Task Description

The ACTI shared task comprises two subtasks, which we describe below.

A: Conspiratorial Content Classification. The first subtask is determining whether a Telegram post is conspiratorial. We consider conspiratorial texts as those that either: (i) express the belief that influential people create major events (e.g., COVID-19) to protect their interests or (ii) interpret events in a way that supports the narrative of a conspiracy theory.

Note that this definition of "conspiratorial" is broad, as *texts* may be defined as conspiratorial if they undermine commonly accepted views on societal issues. For example, the text "il cancro femminista sta prendendo piene" should be classified as conspiratorial, as it subtly supports a broader theory claiming that women's rights are destroying the stability of Western societies.

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT [1]
russog@ethz.ch (G. Russo); niklas.stoehr@inf.ethz.ch (N. Stoehr); manoel.hortaribeiro@epfl.ch (M. H. Ribeiro)
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper36.pdf

B: Conspiracy Category Classification. The second subtask is determining which conspiracy theory a post belongs to. In particular, we consider four possible conspiracy theories.

- **COVID-19**: Text concerning vaccine production, 5G, and non-pharmacological interventions as a tool of control over people. Texts denying the pandemic was a real event or minimizing its importance.
- **QAnon**: Texts associated with the QAnon theory. According to QAnon, a group of Satanic cannibalist sex abusers conspired against former U.S. President Donald Trump during his term in office. This theory extended far over its original scope embodying other beliefs that support (among the others) the idea that women are enemies (hate against women) and that a powerful elite (led by public figures like Pope Francis, Queen Elizabeth, and Hillary Clinton) is trying to organize a New World Order.
- **Flat-Earth**: Texts associated with the claim that the earth is flat and that influential organizations hide this fact from laypeople. Usually, the flat-earth conspiracy theory is supported by pseudo-scientific evidence.
- **Pro-Russia** : Texts associated with conspiratorial beliefs promoting Russian interests, e.g., that nazists control Ukraine's governments and army.

3. Data Collection

To gather the necessary data for the ACTI task, we employ a customized web crawler using the *Selenium* and *BeautifulSoup* libraries in Python. Our web crawler targets specific sources known for hosting conspiratorial content on the Telegram platform.

Specifically, we focus on a selection of Telegram channels that gained notoriety for promoting far-right ideologies and disseminating conspiracy theories. The channels we collect data from include: *Qlobal-Change Italia, Basta Dittatura, Studi Scientifici Vaccini, Terra Piatta,* and *Dentro La Notizia.* For example, the channel "Basta Dittatura" has been actively involved in various events, including the siege of a trade union headquarters, indicating its strong affiliation with conspiratorial movements.

Our data collection process spanned from January 1, 2020, to June 30, 2020, during which we capture and retain comments written in Italian. To ensure sufficient text for analysis, we filtered out comments with less than ten words. We gathered a dataset comprising $25,612$ posts extracted from these five Telegram channels. We summarize statistics about our dataset in fig. 1

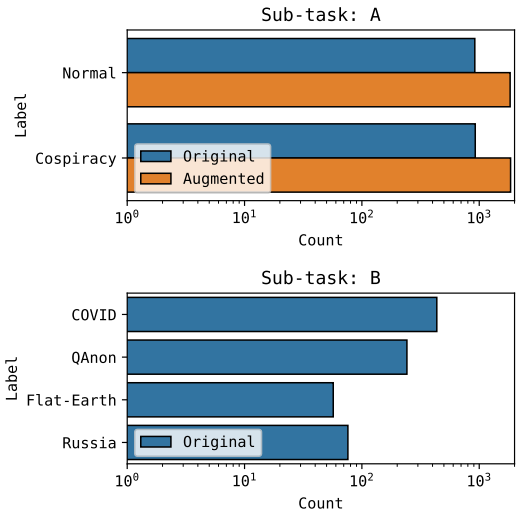


Figure 1: Distribution of labels for Subtask A and Subtask B.

3.1. Annotation Process

The data collection process for our study on conspiratorial content in online channels involved several steps to ensure the accuracy and relevance of the collected data. One of the main challenges we encountered was the presence of non-conspiratorial content within the channels. While some comments discussed conspiratorial topics, others contained valid points or critiques regarding conspiratorial perspectives. Additionally, some comments were deemed meaningless and needed to be filtered out to maintain the integrity of the dataset.

To address this, we employed two human annotators who were responsible for labeling the comments according to three categories: "Not Relevant," "Non-Conspiratorial," and "Conspiratorial." The "Not Relevant" label was assigned to comments that did not contribute to the discussion, while the "Non-Conspiratorial" label was used for comments that did not involve conspiratorial content. The "Conspiratorial" label indicated comments that contained or supported conspiratorial discussions. For the comments labeled as "Conspiratorial," we further categorized them into four subcategories: "QAnon", "Covid19", "Russia", and "Flat-Earth". These subcategories allowed us to analyze specific conspiracy theories in greater detail. The definitions of conspiratorial content are based on established studies in the field [19, 20], ensuring consistency and clarity in our annotation process.

To assess the agreement between the annotators, we calculated inter-annotator agreement rates using Cohen's κ coefficient. The two annotators achieved high agree-

ment levels, with a Cohen's κ of 0.93 for the first task and 0.86 for the second task, demonstrating the reliability of the annotation process. To maintain data integrity, we excluded comments that did not receive the same classification from both annotators. Additionally, comments labeled "Not Relevant" were discarded from the dataset to focus solely on relevant conspiratorial content.

Our data collection process yielded 2,301 comments for the first subtask and 1,110 comments for the second subtask. This resulted in a curated dataset that provides a solid foundation for research on conspiratorial content in online discussions.

4. Evaluation Measures

We chose different evaluation metrics for subtasks A and B because of the distribution of the labels provided by the annotators. In particular

A: Conspiratorial Content Classification. The systems submitted by participants are evaluated using the standard accuracy measures and ranked accordingly.

B: Conspiracy Category Classification. Given the class imbalance for the four types of conspiracy theories we identified, we opt for using as a metric the F1-Score. For a multi-class classification problem, we calculate the F1-score per class in a one-vs-rest manner. We rate each class separately, computing the F1-score for each conspiracy theory in our dataset. To obtain a single score, we then average the per-class F1-scores.

Baselines. We follow the same methodological approach to provide a baseline for both subtasks. Specifically, the baselines for subtasks A and B are a Random Forest trained on a bag-of-words representation of the comments. In particular, we trained the random forest with 500 estimators and validated it using a five-fold cross-validation. These baselines achieve 0.63 accuracy for the first and 0.68 for the second subtask, respectively.

5. Results

A total of fifteen teams submitted from seven institutions participated in the two tasks. Specifically, eight teams submitted for the conspiracy content classification and seven for the conspiracy category classification. In total, we obtain 81 submissions. In Tables 1 and 2, we show the results for both submissions.

Rank	Team Name	Score
1	UPB	0.85712
2	extremITA	0.85647
3	HFI	0.84469
4	Flavio Giobergia	0.83709
5	Michael Vitali	0.82297
6	Giacomo Cignoni	0.82284
7	sCambiaMenti	0.79182
8	Mario Graff	0.78207

Table 1
Conspiratorial Content Classification: Ranking of the eight teams joining the task. The best performing approach was obtained via Contrastive Training

5.1. Conspiratorial Content Classification

Table 1 reports the results of the Conspiratorial Content Classification subtask, which received 40 submissions. The "UPB" team from the University Politehnica of Bucharest achieves the highest accuracy of 0.85 with five submissions. Their methodology consists of an Italian language Sentence Transformer model trained it using contrastive learning. Due to imbalanced data, the participants integrated a data augmentation step in their classification pipeline. Specifically, their methodology generates synthetic data via a Large Language Model (LLM). These synthetic data are then used for training the model. Figure 2 provides an overview of their methodology. The second best-performing team submitted a LLM-based model as well. Specifically, the participants tested extremIT5 (an encoder-decoder model) and extremITLLaMA (an instruction-tuned Decoder-only Large Language Model) designed for handling Italian instructions. While LLM-based approaches performed best, other participants developed methods based on transformers and ensembles, which achieved an accuracy of over 0.80.

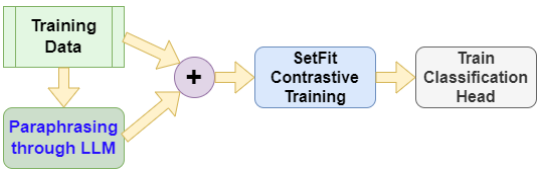


Figure 2: Overview of the methodology used by the 'UPB' team from the University Politehnica of Bucharest, who achieved the highest accuracy of 0.85 with five submissions

5.2. Conspiracy Category Classification

Table 2 reports the results of the Conspiracy Category Classification task, which received 41 submissions in total. Once again, the "UPB" team from the University Politehnica of Bucharest achieved the highest F1-score

Rank	Team Name	Score
1	UPB	0.91225
2	Michael Vitali	0.89826
3	HFI	0.89476
4	Giacomo Cignoni	0.88534
5	extremITA	0.85562
6	Flavio Giobergia	0.83600
7	sCambiaMenti	0.67507

Table 2
Conspiratorial Category Classification: Ranking of the eight teams joining the task. The best-performing approach was obtained via Contrastive Training

(0.91). Interestingly, the data augmentation process used for subtask A did increase the model performance. Indeed, the participants submitted the same transformer-based model trained with contrastive learning, excluding the data augmentation block. The second-best performing team from Tor Vergata University (Michael Vitali) achieved an F1-Score of 0.89. They fine-tuned two BERT models, one in Italian and one multilingual, and combined them in an ensemble. Numerous teams performed well in this task, achieving F1-scores beyond 0.80. Only one participant obtained a result slightly inferior to the provided baseline.

6. Discussion

A comprehensive analysis of the submitted systems reveals that most participants opted for LLMs-based models. Within this context, we emphasize two distinct approaches the participants employ: (i) prompting and (ii) data augmentation. Upon thorough analysis, we find that while prompting does result in positive outcomes, the predictive capabilities of zero-shot LLMs are still inferior to systems that have been finetuned for a specific task.

6.1. Prompting Large Language Models

Prompting consists of providing information to a trained model to predict output labels for a task. It is a task-agnostic approach, making it versatile and widely applicable [21]. This is achieved through concise instructions, referred to as prompts, which guide the model's behavior.

The power and flexibility of prompting LLMs are well exemplified by team ExtremITA's approach: adopting a Large Language Model (LLM) to address all EVALITA tasks simultaneously. For the ACTI task, ExtremITA is prompted with simple questions such as *"Does this text talk about a conspiracy? Answer yes or no"* and *"Which conspiracy theory is discussed in this text: Covid, QAnon, Flat Earth, or Russia?"* This approach ranks second in subtask A of ACTI, achieving a score of 0.86 F1-score. However, it significantly drops in performance in subtask B, ranking fifth with a score of 0.85 F1-score.

In other EVALITA tasks, ExtremITA showed significant variability in predictive capacity. It ranked first in eight out of twenty-five tasks, but in the remaining subtasks, it performed poorly, ranking between fifth and eleventh. These results confirm LLMs' high potential and applicability in real-world scenarios. However, the high variability of results shows that LLMs need help improving over models fine-tuned on specific tasks. Future research should focus on refining prompting techniques to improve the predictive capacity of LLMs at the single-task level.

6.2. Augmenting Data with LLMs

The winning team of subtasks A and B ("UPB") used an approach based on data augmentation via Large Language Models and the training of sentence transformers with contrastive learning. This approach tackles the challenge of the acquisition of conspiratorial data. Indeed, collecting and labeling conspiratorial data requires substantial efforts by domain specialists. This approach tested the possibility of leveraging LLMs to generate synthetic data and use it to train systems for automatically detecting conspiratorial content based. However, it is essential to note that validating the quality of data generated by LLM is an open issue within the NLP community. While LLMs can effectively produce synthetic content, assessing its authenticity and alignment with real-world conspiratorial beliefs is crucial. The lower performance of the model augmented with synthetic data suggests that the quality of the generated data drastically impacts the overall model performance. Therefore, human evaluation is mandatory to evaluate the effectiveness of these approaches.

7. Conclusion

A recent position paper [22] asks whether EVALITA has reached its end in light of the increasing use of LLMs. However, based on the outcomes presented in this report, it becomes evident that the answer remains negative. The challenges posed by EVALITA tasks persist as a crucial asset in comprehending and advancing language resources and tools specifically for the Italian language. This fact is exemplified by transformer-based models' differing rankings, demonstrating the evaluation campaign's diversity and significance. However, the performance achieved by LLMs is undoubtedly pushing the limits of some tasks, especially text classification tasks. In conclusion, while LLMs have shown great potential, EVALITA remains an essential platform for improving language tools for the Italian language.

Acknowledgments

We thank the data annotators for their careful and valuable work. Niklas Stoehr acknowledges funding from the Swiss Data Science Center (SDSC) fellowship.

References

[1] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[2] J. Roisman, The rhetoric of conspiracy in ancient Athens, Univ of California Press, 2006.

[3] Macey, Richard, One giant blunder for mankind: how nasa lost moon pictures, 2006.

[4] J.-W. Van Prooijen, K. M. Douglas, Conspiracy theories as part of history: The role of societal crisis situations, Memory studies 10 (2017) 323–333.

[5] K. M. Douglas, J. E. Uscinski, R. M. Sutton, A. Cichocka, T. Nefes, C. S. Ang, F. Deravi, Understanding conspiracy theories, Political psychology 40 (2019) 3–35.

[6] E. M. Rose, The Murder of William of Norwich: The Origins of the Blood Libel in Medieval Europe, Oxford University Press, 2015.

[7] C. Sunstein, # Republic: Divided democracy in the age of social media, Princeton university press, 2018.

[8] T. Goertzel, Belief in conspiracy theories, Political psychology (1994) 731–742.

[9] Anti-Defamation League, ADL statement on Facebook's decision to finally ban QAnon content from platform, https://www.adl.org/news/press-releases/adl-statement-on-facebooks-decision-to-finally-ban-qanon-content-from-platform, 2020.

[10] BuzzFeed News, 'the rioters who took over the capitol have been planning online in the open for weeks', 2021. URL: https://www.buzzfeednews.com/article/janelytvynenko/trump-rioters-planned-online.

[11] B. Collins, B. Zadrozny, Facebook bans qanon across its platforms, https://www.nbcnews.com/tech/tech-news/facebook-bans-qanon-across-its-platforms-n1242339, 2020.

[12] N. Puri, E. A. Coomes, H. Haghbayan, K. Gunaratne, Social media and vaccine hesitancy: new updates for the era of covid-19 and globalized infectious diseases, Human vaccines & immunotherapeutics 16 (2020) 2586–2593.

[13] J. Friedrichs, N. Stoehr, G. Formisano, Fear-anger contests: Governmental and populist politics of emotion, Online Social Networks and Media 32 (2022) 100240. URL: https://linkinghub.elsevier.com/retrieve/pii/S2468696422000428. doi:10.1016/j.osnem.2022.100240.

[14] E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, E. Gilbert, You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech, Proceedings of the ACM on Human-Computer Interaction 1 (2017) 1–22.

[15] G. Russo, L. Verginer, M. H. Ribeiro, G. Casiraghi, Spillover of antisocial behavior from fringe platforms: The unintended consequences of community banning, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 17, 2023, pp. 742–753.

[16] S. Zannettou, " i won the election!": An empirical analysis of soft moderation interventions on twitter, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 15, 2021, pp. 865–876.

[17] G. Russo, M. Horta Ribeiro, G. Casiraghi, L. Verginer, Understanding online migration decisions following the banning of radical communities, in: Proceedings of the 15th ACM Web Science Conference 2023, 2023, pp. 251–259.

[18] G. Russo, C. Gote, L. Brandenberger, S. Schlosser, F. Schweitzer, Helping a friend or supporting a cause? disentangling active and passive cosponsorship in the U.S. congress, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2952–2969. URL: https://aclanthology.org/2023.acl-long.166.

[19] C. R. Sunstein, A. Vermeule, Conspiracy theories: Causes and cures, Journal of political philosophy 17 (2009) 202–227.

[20] V. Swami, R. Coles, S. Stieger, J. Pietschnig, A. Furnham, S. Rehim, M. Voracek, Conspiracist ideation in britain and austria: Evidence of a monological belief system and associations between individual psychological differences and real-world and fictitious conspiracy theories, British Journal of Psychology 102 (2011) 443–463.

[21] C. Lefebvre, N. Stoehr, Rethinking the event coding pipeline with prompt entailment, in: arXiv, volume 10.48550, 2022. URL: https://arxiv.org/pdf/2210.05257.pdf.

[22] V. Basile, et al., Is EVALITA done? on the impact of prompting on the italian nlp evaluation campaign, in: CEUR Workshop Proceedings, volume 3287, Debora Nozza, Lucia C. Passaro, Marco Polig-

nano, 2022, pp. 127–140.

Vitali at ACTI – Transformer-based Conspiracy Theory Identification

Michael Vitali[1], Vincenzo Scotti[1] and Mark James Carman[1]

[1]DEIB, Politecnico di Milano, Via Ponzio 34/5, 20133, Milano (MI), Italy

Abstract

English. In this work, we participated in the Automatic Conspiracy Theory Identification (ACTI) competition, which involved two sub-tasks: identifying whether an input text is a conspiracy theory and recognising the specific conspiracy theory it discusses. Our approach involved fine-tuning two BERT models, one in Italian and one multilingual, and combining them in an ensemble. The results were promising, and we achieved a position among the top participants in the challenge. This work contributes to the advancement of automatic conspiracy theory identification and highlights the effectiveness of fine-tuned BERT models in this domain.

Italiano. In questo lavoro, abbiamo partecipato alla competizione di Identificazione Automatica delle Teorie Cospiratorie (Automatic Conspiracy Theory Identification, ACTI), che si compone due sotto-problemi: identificare se un dato testo riguarda una teoria del complotto e riconoscere a quale teoria del complotto in particolare si fa riferimento nel testo. Il nostro approccio prevedeva l'adattamento di due modelli BERT, uno in italiano e uno multilingue, e la loro combinazione in un ensemble. I risultati sono stati promettenti e abbiamo raggiunto una posizione tra i primi partecipanti nella sfida. Questo lavoro contribuisce allo sviluppo dell'identificazione automatica delle teorie del complotto e mette in evidenza l'efficacia dei modelli BERT adattati in questo ambito.

Keywords

Natural Language Processing, Transformer Network, BERT, Ensemble, Conspiracy Theory Identification

1. Introduction

The *Automatic Conspiracy Theory Identification* (ACTI) task [1], organised as part of the *Evaluation of Natural Language Processing* (NLP) *and Speech Tools for Italian* (EVALITA) campaign[1] [2], tackles the problem of automating the identification and classification of conspiracy theories. This task arose from the need to address the growing concern surrounding the spread of conspiracy theories on various online platforms, including social media. In the age of *Large Language Models* (LLMs) [3] and widespread disinformation on social media, the ability to track and recognise these theories automatically is highly relevant [4]. In fact, the dissemination of conspiracy theories has the potential to manipulate public opinion, disrupting societal harmony and trust in institutions.

Moreover, while mainstream platforms apply community-level moderation policies, there is a strong ongoing discussion regarding the effectiveness of these methods, as highlighted in recent works [5, 6]. This discussion points out the need for the development of ad-hoc methods to identify rapidly changing and evolving content, such as conspiracy theories. These methods aim to complement and enhance the existing moderation strategies to better address the challenges posed by the proliferation of conspiracy theories in online spaces.

NLP and Deep Learning techniques offer promising solutions that can ease the analysis of large volumes of textual data [7, 8]. Deep Learning-powered NLP models can analyse text data by looking for semantic and lexical cues that can discriminate regular documents from conspiracy theories narratives, and, within the umbrella of conspiracy theories, which one is the subject of the document. In our work, we relied on state-of-the-art transformer neural networks to tackle the problem, combining single classifiers into ensembles to improve the performances as much as possible.

We organised this report into the following sections. In Section 2, we describe the two subtasks composing the ACTI task, the available data, and the preprocessing steps we applied. In Section 3, we present the classification models we built to solve the tasks and we trained such models. In Section 4, we explain the evaluation approach, report our results, and comment on these results. Finally, in Section 5, we summarise our contribution and suggest possible future extensions.

2. Task

In this section, we present the data sets that constitute the two different sub-tasks of the ACTI task (hereafter

EVALITA 2023: Overview of the 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, September 07–08, 2023, Parma, Italy

michael.vitali@mail.polimi.it (M. Vitali); vincenzo.scotti@polimi.it (V. Scotti); mark.carman@polimi.it (M. J. Carman)

0000-0002-8765-604X (V. Scotti); 0000-0001-6575-9737 (M. J. Carman)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).



CEUR Workshop Proceedings (CEUR-WS.org)

[1]Website: https://www.evalita.it

CEUR-WS.org/Vol-3473/paper37.pdf

Table 1
Sub-task data sets main statistics.

Sub-task	Split	No. of samples	Avg. tokens per sample
Original			
A	Train	1842	59.17±73.89
	Test	460	63.67±78.67
B	Train	810	80.09±90.74
	Test	300	77.81±83.95
Pre-processed			
A	Train	3684	51.35±60.17
	Test	460	56.31±66.42
B	Train	810	52.43±65.33
	Test	300	69.34±73.63

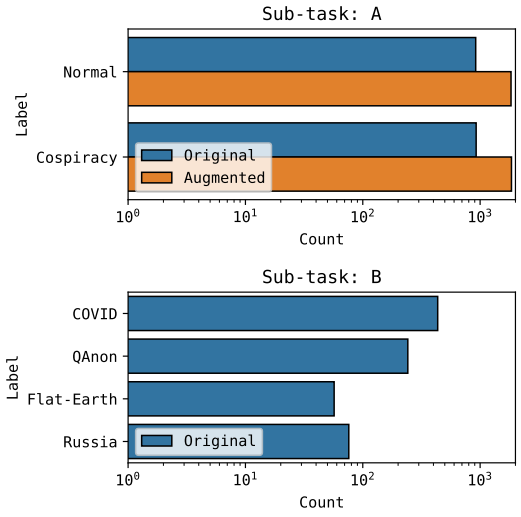


Figure 1: Sub-tasks label distributions.

referred to as *sub-task A* and *sub-task B*). We also describe the pre-processing steps applied to clean and prepare the data sets for training our text-based classifiers.

2.1. Data

The ACTI task comprises two sub-tasks: sub-task A, a binary classification task to determine if a given text piece is about a conspiracy theory or not, and sub-task B, a multi-class classification task to recognise specific conspiracy theories. Both sub-tasks use separate data sets consisting of Italian text samples. Table 1 provides an overview of the main statistics for the text samples in each corpus, we used *NLTK* [9] tokeniser to compute the number of tokens. Figure 1 illustrates the label distributions.

Sub-task A involves binary classification to identify whether a text sample relates to a conspiracy theory or not. Samples contain noise like emoticons or spelling errors, hence we assumed they had not been pre-processed. Label distribution is well balanced between the two classes, as can be seen in the top part of Figure 1.

Sub-task B extends sub-task A by introducing a multi-class classification aspect. The goal is to identify if a text pertains to one of the following conspiracy theories: *COVID*, *QAnon*, *Flat-Earth*, and *Russia*. As for sub-task A, samples have not been pre-processed. Differently from sub-task A, the label distribution is unbalanced, COVID and QAnon are more frequent than Flat-Earth and Russia.

2.2. Pre-processing

Text samples in the data sets contain a lot of noise, like emoticons, slang, or spelling errors; thus, we applied some cleaning and pre-processing steps. Moreover, the two data sets do not contain a large number of samples, and sub-task B presents an unbalanced label distribution, introducing the risks of overfitting and learning biased models. To cope with this issue we considered applying data augmentation to the data sets. Initial results on the training set showed that augmentation was relevant to obtain good results on sub-task A, while we did not need to apply it to sub-task B, despite the class unbalance.

To clean the data sets, we employed basic text transformation and regular expressions to:

- Convert all text to lowercase to ensure consistency and reduce the vocabulary size.
- Clean the data by removing noise such as emoticons, slang, and special characters with regular expressions. This step helps to improve the quality of the text samples.
- Clean data by removing specific patterns from the text, including dates, texts between brackets, links, emails, and multiple spaces, using regular expressions.

Additionally, we applied data augmentation to sub-task A, to increase the number of samples. The method of choice was *back-translation*, which involves translating a text sample from the source language to another language and then translating it back. This process preserves the original text's semantics while potentially altering the syntax, generating synthetic samples.

3. Model

In this section, we describe the architecture of the classifiers we built using *Transformer neural networks* [10] and the training process we followed to prepare our models for evaluation.

3.1. Architecture

To solve both ACTI sub-tasks, we used the same Transformer-based classification architecture, changing only the target classes from one task to the other. We explored different *Transformer Encoder* neural networks, namely BERT [11], pre-trained on different data sets. We also explored individual and combined applications of the classifiers. We visualise the single-model and ensemble-model pipelines in Figure 2.

Each BERT-based classifier takes as input a sequence of tokens, extracted from the pre-processed text. The sequence starts with a classification token `[CLS]` and is concluded by an end-of-sequence token `[SEP]`, introduced during the tokenisation process. To classify the input piece of text, we retrieve the *contextual embedding* computed by the transformer *hidden layers* in correspondence of the `[CLS]` input token and use it to feed a linear classifier. The final classification layer outputs the probability distribution of the input piece of text to belong to one of the possible classes. We reported the entire process in Figure 2a.

To improve the classification results and take the best from the trained models, we considered also creating an *ensemble* [12, Chapter 16]. For each task, we aggregated the predictions of the individual models. To aggregate the predictions, we froze the fine-tuned classifiers and learned a separate *Logistic Regression* classifier on top of the Transformer models. The Logistic Regression classifier takes as input the probability distributions predicted by individual models and compute a new output probability combining the previous. The entire ensemble pipeline is represented in Figure 2b.

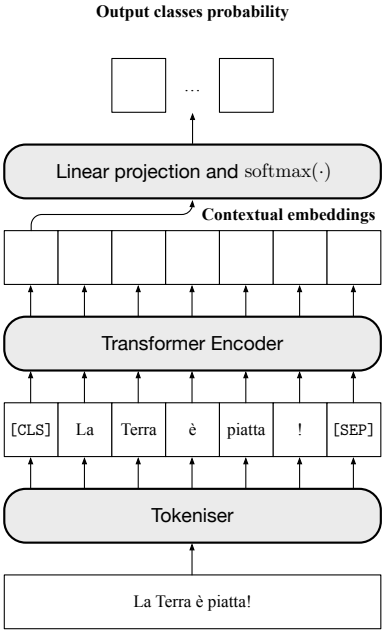
3.2. Training

To effectively train our models, we adopted *5-fold cross-validation* to find the best hyperparameters for each of the considered models and each task. We preferred this approach to the usual train-validation split to make the best out of the available data. Given the best hyperparameters combination, we retrained the model on the entire training data set.

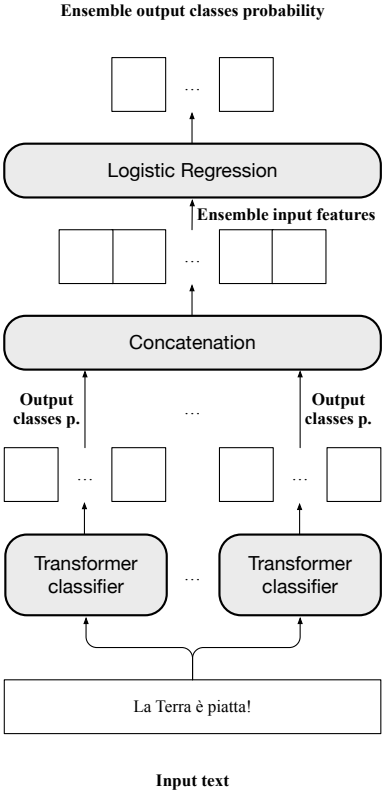
We fine-tuned two variants of BERT base (110M parameters):

- Italian BERT (uncased)[2], pre-trained on Italian.

[2]Model card:
https://huggingface.co/dbmdz/bert-base-italian-xxl-uncased



(a) Single Transformer classifier.



(b) Ensemble classifier.

Figure 2: BERT-based classifiers architecture.

- Multilingual BERT (uncased)[3]; pre-trained on several languages, including Italian.

We used the implementations available in the *Transformers* library from *Hugging Face* [13]. Each configuration was trained using the *Adam* optimiser [14], a linear learning rate schedule and a batch size of 8.

For each of these models, during cross-validation, we varied the learning rate and the number of epochs. Additionally, we explored regularisation: we evaluated models with and without *dynamic masking*. We varied the learning rate in $\{1\times10^{-5},2\times10^{-5},3\times10^{-5},5\times10^{-5}\}$, and the number of epochs in $\{2,3,4\}$. Dynamic masking applies the same kind of masking BERT uses during pre-training, randomly corrupting the input sequence by masking the tokens.

We adopted 5-fold cross-validation with the ensemble as well. Referring to the implementation of Logistic Regression available in the *Scikit-Learn* library [15], we explored values for the following hyperparameters: regularisation strength (L_2 regularisation), number of iterations, and solver. We varied the inverse of the regularisation strength in $\{10^{-3},10^{-2},10^{-1},1,10,10^{2},10^{3}\}$, the maximum number of iterations in $\{20,50,100,200,500,1000\}$, and we tried all solvers apart from the Newton-Cholesky one. Additionally, we weighed the classes with a weight inversely proportional to class frequencies, to obtain a more balanced classifier.

4. Results

In this section, we explain how we evaluated the models proposed for each sub-task, present the results obtained on each task, and provide comments on these results. In both cases, we focus in the results of the ensembles, since they peeform better than individual models in both cases.

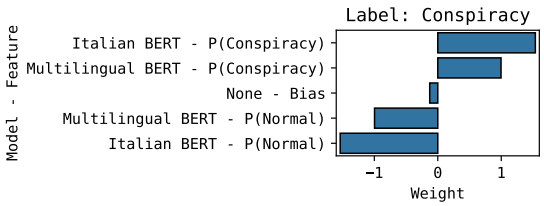
4.1. Evaluation

We evaluated the classification models using the F_1 score. For the multitask settings, we computed the macro average of the scores to account for potential class distribution imbalances.

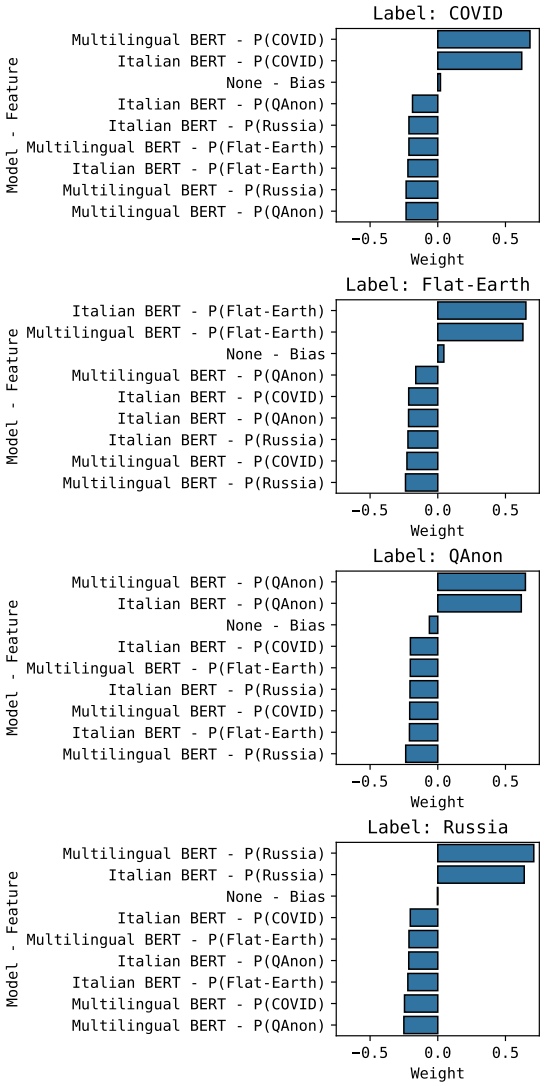
We reported the F_1 scores on the test set in table Table 2. In addition to the results of the submitted models, we included some additional scores for comparison and to provide further insight into the results. The F_1 scores are computed on 70% of the test data for sub-task A 50% of the test data for sub-task B via the *Kaggle* platform[4] (which hosted the competition), as determined by the authors of the ACTI task for the private leaderboard.

[3]Model card:
https://huggingface.co/bert-base-multilingual-uncased

[4]Website: https://www.kaggle.com



(a) Sub-task A.



(b) Sub-task B.

Figure 3: Feature relevance in the ensembles. Relevance scores are given by the weights learned by the Logistic Regression classifier.

Table 2
Ensemble results on ACTI task. Scores are from the private Kaggle evaluation on 70% of the test data. Baselines were provided by task organisers via Kaggle.

Sub-task	Augmentation	Dynamic Masking	$\mathbf{F}_1$
Submitted			
A	✓	✓	**82.30**
B		✓	**89.83**
Other			
A	✓		77.04
A			81.67
B			87.67
Baselines			
A	n.a.	n.a.	51.07
B	n.a.	n.a.	68.37

Additionally, to get better insights on the behaviour of the two Transformer-based classifiers on the sub-tasks, we analysed the weights learned by the Logistic Regression during the training of the ensemble. The higher the weight, the stronger the contribution of the probability predicted by a classifier to the prediction and, thus, the stronger the relevance of that classifier in the ensemble. To this end, we visualised all the weights of the two Logistic Regression models in Figure 3.

4.2. Sub-task A

For Sub-task A, the ensemble model achieved a test F_1 score of 82.30% (see Table 2). This result highlights the effectiveness of combining the predictions from individual models to improve overall performance.

The best model configuration for sub-task A, involved fine-tuning with the following hyperparameters. Italian BERT: learning rate of 2×10^{-5}, number of epochs of 4, and dynamic masking enabled. Multilingual BERT: learning rate of 2×10^{-5}, number of epochs of 4, and dynamic masking enabled. Logistic Regression (ensemble): inverse of the regularisation strength of 10^{-2}, maximum number of iterations of 20, *Newton-CG* solver.

Comparing it with other configurations, we observe that the ensemble model outperformed other approaches, such as using augmentation alone (77.04%) or not applying any augmentation nor masking (81.67%). Furthermore, compared to the provided baseline F_1 score of 51.07%, the proposed ensemble model shows a substantial improvement, highlighting the effectiveness of our approach.

Concerning feature relevance analysis in the ensemble, from Figure 3a, we can see that the ensemble gives a higher weight to the prediction of the Italian BERT model, rather than the multilingual one. This hints that for this specific task of detecting whether the text concerns a conspiracy theory or not, having a language-specific model may be the better solution. However, the ensemble improves over both single models, thus the multilingual model is contributing to the correct classification as well.

4.3. Sub-task B

For Sub-task B, the ensemble model achieved a test accuracy score of 89.83% (see Table 2). This result highlights again the effectiveness of the ensemble approach in capturing task-specific patterns and making accurate predictions.

The best model configuration for sub-task B, involved fine-tuning with the following hyperparameters. Italian BERT: learning rate of 3×10^{-5}, number of epochs of 2, and dynamic masking enabled. Multilingual BERT: learning rate of 3×10^{-5}, number of epochs of 2, and dynamic masking enabled. Logistic Regression (ensemble): inverse of the regularisation strength of 10^{-3}, maximum number of iterations of 20, *Newton-CG* solver.

Comparing it with other configurations, we observe that the ensemble model outperformed the configuration without dynamic masking, which achieved an accuracy score of 87.67%. This indicates that dynamic masking played a crucial role in improving the model's performance. When comparing the ensemble model's accuracy score with the provided baseline accuracy score of 68.37%, we observe a significant performance boost, underscoring the effectiveness of our approach.

Concerning feature relevance analysis in the ensemble, from Figure 3b), we can see that, differently from sub-task A, here both models contribute equally to the prediction. In fact, the values of the weights associated with the same input probability and the same output class models are close for the different models.

5. Discussion

In this report, we described our approach to training Transformer-based classification models for conspiracy theory identification. We trained and evaluated our models on the two sub-tasks of the ACTI data set, an Italian benchmark for conspiracy theory identification. The first task involved binary text classification to determine whether a piece of text is about a conspiracy theory or not, while the second task focused on multi-class classification to identify the specific conspiracy theory referenced in a piece of text.

Given the limited resources available, including the data set size and computational power, we were unable to explore all possible avenues. Additionally, the availability of pre-trained Italian Language Models is also limited. Most Italian models are part of multilingual models rather than dedicated Italian models, and the available Italian-only models are smaller if compared to English ones (for example). However, this provided us with the opportunity to train potentially multilingual models for conspiracy theory identification, although we did not test this approach on other languages in our current study.

The results obtained from our models are promising. For Sub-task A, our ensemble model achieved a test F_1 score of 82.30%, outperforming both the other configurations we explored and the provided baseline F_1 score of 51.07%. Regarding Sub-task B, our ensemble model achieved a test F_1 score of 89.83%, surpassing the other configurations we tested and the provided baseline F_1 score of 68.37%. This highlights the effectiveness of combining the predictions from individual models to improve overall performance on this task.

Moving forward, we aim to explore the application of end-to-end text generation models for conspiracy theory identification. Current research suggests that LLMs can be effectively employed for text classification tasks by concatenating the text to classify with a question asking for the class and triggering text generation. We plan to leverage one of these multilingual LLMs with a combination of *prompting* and *in-context learning*, enabling *zero-shot* to *few-shots* learning.

Overall, our study contributes to the understanding of conspiracy theory identification using Transformer-based models. The achieved results show the potential of these models in accurately classifying conspiracy-related texts, and future investigations can explore additional approaches to further enhance performance.

References

[1] G. Russo, N. Stoehr, M. H. Ribeiro, Acti at evalita 2023: Overview of the conspiracy theory identification task, arXiv preprint arXiv:2307.06954 (2023).

[2] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[3] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J. Nie, J. Wen, A survey of large language models, CoRR abs/2303.18223 (2023). URL: https://doi.org/10.48550/arXiv.2303.18223. doi:10.48550/arXiv.2303.18223. arXiv:2303.18223.

[4] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh, Z. Kenton, S. Brown, W. Hawkins, T. Stepleton, C. Biles, A. Birhane, J. Haas, L. Rimell, L. A. Hendricks, W. Isaac, S. Legassick, G. Irving, I. Gabriel, Ethical and social risks of harm from language models, CoRR abs/2112.04359 (2021). URL: https://arxiv.org/abs/2112.04359. arXiv:2112.04359.

[5] G. Russo, M. Horta Ribeiro, G. Casiraghi, L. Verginer, Understanding online migration decisions following the banning of radical communities, in: Proceedings of the 15th ACM Web Science Conference 2023, WebSci '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 251–259. URL: https://doi.org/10.1145/3578503.3583608. doi:10.1145/3578503.3583608.

[6] G. Russo, L. Verginer, M. H. Ribeiro, G. Casiraghi, Spillover of antisocial behavior from fringe platforms: The unintended consequences of community banning, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 17, 2023, pp. 742–753.

[7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[8] G. Russo, C. Gote, L. Brandenberger, S. Schlosser, F. Schweitzer, Helping a friend or supporting a cause? disentangling active and passive cosponsorship in the U.S. congress, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2952–2969. URL: https://aclanthology.org/2023.acl-long.166.

[9] S. Bird, E. Klein, E. Loper, Natural Language Processing with Python, O'Reilly, 2009. URL: http://www.oreilly.de/catalog/9780596516499/index.html.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017,

Long Beach, CA, USA, 2017, pp. 5998–6008. URL: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[11] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186. URL: https://doi.org/10.18653/v1/n19-1423. doi:10.18653/v1/n19-1423.

[12] T. Hastie, R. Tibshirani, J. H. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition, Springer Series in Statistics, Springer, 2009. URL: https://doi.org/10.1007/978-0-387-84858-7. doi:10.1007/978-0-387-84858-7.

[13] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-art natural language processing, in: Q. Liu, D. Schlangen (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020, Association for Computational Linguistics, 2020, pp. 38–45. URL: https://doi.org/10.18653/v1/2020.emnlp-demos.6. doi:10.18653/v1/2020.emnlp-demos.6.

[14] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. URL: http://arxiv.org/abs/1412.6980.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. VanderPlas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830. URL: https://dl.acm.org/doi/10.5555/1953048.2078195. doi:10.5555/1953048.2078195.

A. Source Code

The source code developed for the challenge is available via *GitHub* at the following link: https://github.com/MichaelVitali/Evalita2023.

UPB at ACTI: Detecting Conspiracies using fine tuned Sentence Transformers

Andrei Paraschiv[1], Mihai Dascalu[1]

[1]University Politehnica of Bucharest 313 Splaiul Independetei, Bucharest, Romania

Abstract

English. Conspiracy theories have become a prominent and concerning aspect of online discourse, posing challenges to information integrity and societal trust. As such, we address conspiracy theory detection as proposed by the ACTI @ EVALITA 2023 shared task. The combination of pre-trained sentence Transformer models and data augmentation techniques enabled us to secure first place in the final leaderboard of both sub-tasks. Our methodology attained F1 scores of 85.71% in the binary classification and 91.23% for the fine-grained conspiracy topic classification, surpassing other competing systems.

Keywords

Conspiracy Theory, Content Moderation, Large Language Models, Computational Social Science

1. Introduction

Conspiracy theories distort the shared understanding of reality and erode trust in crucial democratic institutions. By substituting reliable, evidence-based information with dubious, implausible, or blatantly false claims, these theories foster a climate of disagreement regarding facts and give undue weight to personal opinions and anecdotal evidence over established facts and scientifically validated theories. Aaronovitch [2] defines conspiracy theories as 'the attribution of deliberate agency to something more likely to be accidental or unintended; therefore, it is the unnecessary assumption of conspiracy when other explanations are more probable.' Due to the rapid spread of information across the internet, coupled with the alarming speed at which false information can proliferate [3], we find ourselves amidst what some have dubbed a "golden age" of conspiracy theories [4]. Being a distinct form of misinformation, conspiracy theories exhibit unique characteristics. Brotherton et al. [5] identified five key attributes commonly found in modern conspiracy theories: government malfeasance, extraterrestrial cover-up, malevolent global conspiracies, personal well-being, and information control.

While embracing conspiracy theories can give individuals a sense of reclaiming power or accessing hidden knowledge, these beliefs can sometimes have negative and dangerous consequences. One recent example is the violent insurrection on the US Capitol on 6 January 2021 driven by conspiracy theories surrounding QAnon and election fraud [6]. Additionally, these theories can serve as powerful tools in the hands of nefarious groups, politicians, or state actors who exploit susceptible communities, manipulating them into taking or endorsing actions that can result in significant and dramatic social repercussions [7, 8].

Building upon the importance of addressing conspiracy theories, efforts have been made to research and develop automated methods for detecting conspiratorial content on various platforms and languages. For instance, as part of the EVALITA 2023 workshop, the organizers of the ACTI shared task introduced a novel approach: the automatic identification of conspiratorial content in Italian language Telegram messages. This initiative aimed to enhance our ability to quickly recognize and respond to conspiracy theories, enabling the promotion of critical thinking and media literacy by providing reliable sources and encouraging evidence-based discourse. Leveraging such advancements can effectively limit the influence of conspiracy theories while fostering a more informed and resilient society.

This paper presents our contribution to the ACTI @ EVALITA 2023 shared task [9]. We focused on employing the power of pretrained Italian language sentence Transformers. To further enhance the performance and address potential biases, we employed Large Language Models (LLMs) to augment the training data, resulting in a more balanced and comprehensive training set. This combination of leveraging pre-trained models and data augmentation techniques formed the foundation of our methodology, enabling us to achieve first place in the final leaderboard of both sub-tasks with F1 scores of 85.71% and respectively 91.23%.

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT [1]

andrei.paraschiv74@upb.ro (A. Paraschiv);
mihai.dascalu@upb.ro (M. Dascalu)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper38.pdf

2. Related Work

Recently, online platforms have often banned—entirely deactivated—communities that breached their increasingly comprehensive guidelines. In 2020 alone, Reddit banned around 2,000 subreddits (the name a community receives on the platform) associated with hate speech. Similarly, Facebook banned 1,500 pages and groups related to the QAnon conspiracy theory [10]. While these decisions are met with enthusiasm [e.g., see Anti-Defamation League [11]], the efficacy of "deplatforming" these online communities has been questioned [12, 13]. When mainstream platforms ban entire communities for their offensive rhetoric, users often migrate to alternative *fringe platforms*, sometimes created exclusively to host the banned community [14, 15]. Banning, in that context, would not only strengthen the infrastructure hosting these fringe platforms [12] but allow these communities to become more toxic elsewhere [16]. In order to improve the efficacy of such moderation policies identifying and tracking the propagation of problematic content like conspiracy theories is crucial. For example the Zika virus outbreak in 2016, coupled with the influence of social networks and the declaration of a public health emergency by the WHO, showed the harm the dissemination of conspiracy theories can generated [17, 18].

The COVID-19 pandemic had a profound impact, emphasizing the dangers associated with the proliferation of conspiracy theories. These theories encompassed a wide range of topics, including the virus's origin, its spread, the role of 5G networks, and the efficacy and safety of vaccines. With COVID-related lockdowns in place, people became more reliant on social networking platforms such as Twitter, Facebook, and Instagram, which increased their exposure to disinformation and conspiracy theories. MediaEval 2020 [19] focused on a 5G and COVID-19 conspiracy tweets dataset, proposing two shared tasks to address this issue. The first task involved detecting conspiracies based on textual information, while the second task focused on structure-based detection utilizing the retweet graph. Various systems were proposed to tackle these tasks, employing different approaches such as methods relying on Support Vector Machine (SVM) [20], BERT [21], and GNN [22] . In their study, Tyagi and Carley [23] employed an SVM to classify the stance of Twitter users towards climate change conspiracies. Their findings revealed that individuals who expressed disbelief in climate change tend to share a significantly higher number of other types of conspiracy-related messages compared to those who believe in climate change. Furthermore, Amin et al. [24] manually labeled 598 Facebook comments as Covid-19 vaccine conspiracy or neutral and used a BERT-based model in conjunction with Google Perspective API to classify these messages, providing valuable insights into the prevalence of vaccine conspiracy theories on social media platforms.

Tunstall et al. [25] presented a new approach based on Sentence Transformers[26] called SetFit that focused on data-efficient fine-tuning of sentence embeddings, particularly for binary labels. The training of SetFit follows a two-step process. First, it fine-tuned the sentence embeddings in a contrastive manner. This step helped in optimizing the embeddings for the specific classification task. Subsequently, a classification head was trained using finetuned sentence embeddings, enabling effective classification on the training labels. Their approach aimed to enhance the efficiency and performance of fine-tuning sentence embeddings in scenarios with limited data. The efficacy and power of Sentence-Transformers has been shown in multiple tasks spanning from text generation [27, 28] to sentence classification tasks.[29, 30, 31]. These models capture the semantic and contextual information of sentences or paragraphs, enabling nuanced representations of textual data. Leveraging such models, Bates and Gurevych [32] used SetFit to propose LAGONN, a hate speech and toxic messages classification framework for content moderation.

3. Method

3.1. Task Description

The ACTI @ EVALITA 2023 organizers put forth two subtasks for participants to address. The first sub-task [9] involved binary classification, where participants were provided with a dataset consisting of 1,842 training samples and 460 test samples. The objective was to classify messages as either conspiratorial or non-conspiratorial. The second sub-task focused on fine-grained conspiracy topic classification. Participants were required to classify messages into one of four specific conspiracy topic classes: Covid, QAnon, Flat-Earth, or Russia-conspiracy. A training set of 810 records was provided for this subtask, while the evaluation test set contained 300 samples. Table 1 shows the class distribution for both sub-tasks.

	Classes	**Count**
Sub-Task A	Non Conspiratorial	917
	Conspiratorial	925
Sub-Task B	Covid	435
	QAnon	242
	Flat-Earth	76
	Russian	57

Table 1
ACTI Dataset distribution for the training sets on Sub-task A and B.

The macro F1 score was adopted as a criterion to evaluate the two sub-tasks. During the competition, 30% of the test dataset was immediately evaluated on the Public Leaderboard, giving participants an initial indication of their model's performance. However, the final evaluation was conducted on the remaining 70% of private entries. These final evaluation scores were then used to compile the Private Leaderboard made public after the conclusion of the competition.

3.2. Sentence Transformer and Data Augmentation

We considered an Italian language Sentence Transformer model for our submissions and trained contrastive with SetFit[1] as described by Tunstall et al. [25]. Since the training dataset is highly imbalanced between the conspiratorial classes (see in Table 1), we integrated a data augmentation step in our classification pipeline, as seen in Figure 1.

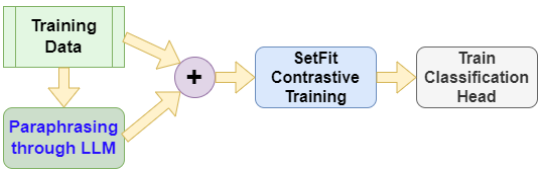


Figure 1: End-to-End training Pipeline.

In the data augmentation step, we used an LLM to create paraphrases for our training data using the prompt "riformulare questo testo: [*comment_text*]" and different seeds to create variations of the answers. In our experiments, we used "text-davinci-003" from the GPT-3 family[2] and the mT5 model finetuned on Italian language paraphrases[3]. We set a high temperature (t=0.9) for the LLMs to ensure diverse text generation. The distribution for the augmented dataset is shown in Table 2.

	Classes	Count
Sub-Task A	Non Conspiratorial	1,822
	Conspiratorial	2,524
Sub-Task B	Covid	779
	QAnon	672
	Flat-Earth	362
	Russian	322

Table 2
Class distribution on the augmented training sets used for Sub-task A and B

[1]https://github.com/huggingface/setfit
[2]https://platform.openai.com/docs/models
[3]https://huggingface.co/aiknowyou/mt5-base-it-paraphraser

Sentence-Transformers are pretrained Transformer models finetuned in a Siamese network, such that semantically similar sentences or paragraphs are projected near each other in the embedding space; in contrast, the distance in the embedding space is maximized for sentence pairs that are different. In our experiments, we used several Italian pretrained Sentence Transformers from the Huggingface Hub[4], as mentioned in Table 3. The first step in the SetFit training process involves generating positive and negative triplets. Positive triplets consist of sentences from the same class, while negative triplets contain sentences from different classes. The training data is expanded by including positive and negative triplets, providing a more comprehensive and diverse training set. The Sentence Transformer captures the contextual and semantic information of the messages, providing a powerful feature representation. In the second step, a fully connected classification head is trained on top of the Sentence-Transformer to distinguish between the available classes.

4. Results

Besides experimenting with different pre-trained models, as shown in Table 3, we also performed grid search tuning with several key hyper-parameters, namely the number of iterations, the learning rate, and the number of epochs for training. The number of iterations determined the quantity of generated triplets during training. By adjusting this parameter, we controlled the training data's size, potentially influencing the model's ability to generalize and capture important patterns. We set the maximum sequence length for the tokenizer to 512 for all of our experiments. We withheld 20% of the training data to evaluate the performance of the trained models during the development time.

The best-performing model differed between the sub-tasks. The best-performing model in the binary classification sub-task was based on "efederici/sentence-BERTino". This model was trained on the "text-davinci-003" augmented dataset for 1 epoch. We used 5 iterations and a learning rate of 1e-05. In contrast, the larger "nickprock/sentence-bert-base-italian-xxl-uncased" model performed best for the fine-grained conspiracy topic classification sub-task. We trained this model on the same dataset for 1 epoch. The learning rate used was 1e-05, and the number of iterations was set to 10. This model yielded the best results in both Leaderboards (see Table 4).

We conducted an ablation study after the competition ended to assess the impact of data augmentation. We trained the best-performing models under different

[4]https://huggingface.co/models

Model	Embedding Size
efederici/sentence-BERTino	768
efederici/sentence-bert-base	768
efederici/sentence-BERTino-3-64	64
efederici/mmarco-sentence-BERTino	768
efederici/sentence-it5-base	512
efederici/sentence-it5-small	512
nickprock/sentence-bert-base-italian-uncased	768
nickprock/sentence-bert-base-italian-xxl-uncased	768
aiknowyou/aiky-sentence-bertino	768

Table 3
Sentence-Transformer models considered in our experiments.

	Public Leaderboard	Private Leaderboard
Sub-Task A	85.36%	85.71%
Sub-Task B	87.62%	91.23%

Table 4
Best model performance on the Public and Private learderboard

conditions: a) using the original training data with 20% reserved for development evaluation, b) considering the entire original training data, and c) employing the dataset that was augmented with the mT5 paraphrasing LLM. The results in Tables 5 and 6 show the importance of the augmentation step.

	Public Leaderboard	Private Leaderboard
No augmented data	75.80%	81.29%
No augmented data with development set	79.36%	83.83%
mT5 augmented Data	78.15%	82.25%

Table 5
Ablation Study for Sub-Task A.

	Public Leaderboard	Private Leaderboard
No augmented data	83.32%	93.67%
No augmented data with development set	83.39%	89.67%
mT5 augmented Data	83.24%	87.07%

Table 6
Ablation Study for Sub-Task B.

In the case of sub-task A, the additional data substantially influenced both the Public and Private test results. The augmented dataset led to significant improvements in performance. However, we see a decline in the Private Leaderboard for the fine-grained task results as the amount of data increased, despite the Public Leaderboard performance keeping the same. This performance decline could be attributed to an unusual distribution difference between the Public and Private test rows. Furthermore, the quality of the paraphrases used in the augmentation process played a crucial role in both sub-tasks. The poor performance achieved by the mT5 model suggests that the quality of the generated paraphrases has a notable impact on the overall model performance. Similarly, a drastic decrease in performance was observed for the second sub-task private leaderboard, arguing for the questionable quality of the paraphrases.

5. Conclusion

In this paper, we described our approach addressing the two sub-tasks in the ACTI @ EVALITA 2023 competition. The challenge focuses on automatically detecting conspiratorial Telegram messages and the classification into four conspiracy topics: Covid, QAnon, Flat-Earth, and Russian conspiracies. Through the utilization of text augmentation techniques and the training of Sentence-Transformers with contrastive learning, we developed robust classifiers. Our best models achieved first place in the Private Leaderboard on both tasks with F1 scores of 85.712% in the binary classification and 91.225% for the fine-grained conspiracy topic classification. This paper contributes to the growing body of research on conspiracy theory detection and emphasizes the effectiveness of leveraging pre-trained models and data augmentation techniques. Our results argue the potential of these approaches in addressing the challenges posed by conspiracy theories and their propagation in online platforms.

Acknowledgement

This work was supported by a grant of the Ministry of Research, Innovation, and Digitalization, project CloudPrecis, Contract 344/390020/06.09.2021, MySMIS code:

124812, within POC.

References

[1] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[2] D. Aaronovitch, Voodoo histories: The role of the conspiracy theory in shaping modern history, NY: Riverhead (2010).

[3] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, science 359 (2018) 1146–1151.

[4] H. W. Hanley, D. Kumar, Z. Durumeric, A golden age: Conspiracy theories' relationship with misinformation outlets, news media, and the wider internet, arXiv preprint arXiv:2301.10880 (2023).

[5] R. Brotherton, C. C. French, A. D. Pickering, Measuring belief in conspiracy theories: The generic conspiracist beliefs scale, Frontiers in psychology 4 (2013) 279.

[6] A. Seitz, Mob at u.s. capitol encouraged by online conspiracy theories, The Associated Press (2021). URL: https://apnews.com/article/donald-trump-conspiracy-theories-michael-pence-media-social-media-daba3f5dd16a431abc627a5cfc922b87.

[7] W. Audureau, Why conspiracy theorists and the kremlin echo each other's disinformation, Le Monde (2023). URL: https://www.lemonde.fr/en/les-decodeurs/article/2023/03/02/conspiracy-theorists-the-kremlin-echo-each-other-s-disinformation_6017960_8.html.

[8] I. Yablokov, Russian disinformation finds fertile ground in the west, Nature Human Behaviour 6 (2022) 766–767.

[9] G. Russo, N. Stoehr, M. H. Ribeiro, Acti at evalita 2023: Overview of the conspiracy theory identification task, arXiv preprint arXiv:2307.06954 (2023).

[10] B. Collins, B. Zadrozny, Facebook bans qanon across its platforms, https://www.nbcnews.com/tech/tech-news/facebook-bans-qanon-across-its-platforms-n1242339, 2020.

[11] Anti-Defamation League, ADL statement on Facebook's decision to finally ban QAnon content from platform, https://www.adl.org/news/press-releases/adl-statement-on-facebooks-decision-to-finally-ban-qanon-content-from-platform, 2020.

[12] E. Zuckerman, C. Rajendra-Nicolucci, Deplatforming our way to the alt-tech ecosystem, Knight First Amendment Institute at Columbia University, January 11 (2021).

[13] G. Russo, L. Verginer, M. H. Ribeiro, G. Casiraghi, Spillover of antisocial behavior from fringe platforms: The unintended consequences of community banning, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 17, 2023, pp. 742–753.

[14] C. Dewey, Washington Post — These are the 5 subreddits Reddit banned under its game-changing anti-harassment policy, and why it banned them, https://wapo.st/3AO7pbl, 2016.

[15] G. Russo, M. Horta Ribeiro, G. Casiraghi, L. Verginer, Understanding online migration decisions following the banning of radical communities, in: Proceedings of the 15th ACM Web Science Conference 2023, WebSci '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 251–259. URL: https://doi.org/10.1145/3578503.3583608. doi:10.1145/3578503.3583608.

[16] M. Horta Ribeiro, S. Jhaver, S. Zannettou, J. Blackburn, G. Stringhini, E. De Cristofaro, R. West, Do platform migrations compromise content moderation? evidence from r/the_donald and r/incels, Proceedings of the ACM on Human-Computer Interaction 5 (2021) 1–24.

[17] A. Ghenai, Y. Mejova, Catching zika fever: Application of crowdsourcing and machine learning for tracking health misinformation on twitter, in: 2017 IEEE International Conference on Healthcare Informatics (ICHI), 2017, pp. 518–518. doi:10.1109/ICHI.2017.58.

[18] M. J. Wood, Propagating and debunking conspiracy theories on twitter during the 2015–2016 zika virus outbreak, Cyberpsychology, behavior, and social networking 21 (2018) 485–490.

[19] K. Pogorelov, D. T. Schroeder, L. Burchard, J. Moe, S. Brenner, P. Filkukova, J. Langguth, Fakenews: Corona virus and 5g conspiracy task at mediaeval 2020., in: MediaEval, 2020.

[20] M. Moosleitner, B. Murauer, G. Specht, Detecting conspiracy tweets using support vector machines., in: MediaEval, 2020.

[21] A. Malakhov, A. Patruno, S. Bocconi, Fake news classification with BERT, in: Working Notes Proceedings of the MediaEval 2020 Workshop, Online, 14-15 December 2020, volume 2882 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: https://ceur-ws.org/Vol-2882/paper38.pdf.

[22] A. Paraschiv, G.-E. Zaharia, D.-C. Cercel, M. Dascalu, Graph convolutional networks applied to fakenews: corona virus and 5g conspiracy, UPB Scientific Bulletin, Series C: Electrical Engineering

83 (2021) 71–82.
[23] A. Tyagi, K. M. Carley, Climate change conspiracy theories on social media, arXiv preprint arXiv:2107.03318 (2021).
[24] M. H. Amin, H. Madanu, S. Lavu, H. Mansourifar, D. Alsagheer, W. Shi, Detecting conspiracy theory against covid-19 vaccines, arXiv preprint arXiv:2211.13003 (2022).
[25] L. Tunstall, N. Reimers, U. E. S. Jo, L. Bates, D. Korat, M. Wasserblat, O. Pereg, Efficient few-shot learning without prompts, arXiv preprint arXiv:2209.11055 (2022).
[26] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019).
[27] A. Amin-Nejad, J. Ive, S. Velupillai, Exploring transformer text generation for medical dataset augmentation, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, 2020, pp. 4699–4708.
[28] G. Russo, N. Hollenstein, C. C. Musat, C. Zhang, Control, generate, augment: A scalable framework for multi-attribute text generation, ArXiv abs/2004.14983 (2020).
[29] J. Hong, J. Park, D. Kim, S. Choi, B. Son, J. Kang, Empowering sentence encoders with prompting and label retrieval for zero-shot text classification, 2023. arXiv:2212.10391.
[30] G. Piao, Scholarly text classification with sentence bert and entity embeddings, in: Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2021 Workshops, WSPA, MLMEIN, SDPRA, DARAI, and AI4EPT, Delhi, India, May 11, 2021 Proceedings 25, Springer, 2021, pp. 79–87.
[31] G. Russo, C. Gote, L. Brandenberger, S. Schlosser, F. Schweitzer, Disentangling active and passive cosponsorship in the u.s. congress, ArXiv abs/2205.09674 (2022).
[32] L. Bates, I. Gurevych, Like a good nearest neighbor: Practical content moderation with sentence transformers, arXiv preprint arXiv:2302.08957 (2023).

sCambiaMenti at ACTI: Ensemble model with majority voting for Automatic Conspiracy Detection

Selene Bianco[1], Daniela Salusso[2]

[1]AizoOn Consulting S.r.l.
[2]University of Turin

Abstract

English. In this paper we describe the methodology that we have implemented to solve the subtask A of the Automatic Conspiracy Detection (ACTI) task (EVALITA 2023). We have developed different classifiers and then used a majority voting approach to obtain the final prediction. The implemented classifiers can be distinguished into three different types: Machine Learning models trained on a document-term matrix (Support Vector Machines, Random Forest and Multinomial Näive Bayes), Neural Network models trained on the text sequences (Long Short Term Memory), and Machine Learning models trained on a set of linguistic features derived from the text. While the single models were prone to overfitting, the classification obtained with the voting approach appears to be more stable and showed an adequate performance on the official test set of the contest.

Italian. *In questo articolo descriviamo la metodologia implementata per risolvere il subtask A del task Automatic Conspiracy Detection (ACTI) task (EVALITA 2023). Abbiamo sviluppato diversi classificatori, per poi ottenere una classificazione finale tramite un meccanismo di voto a maggioranza. I classificatori implementati possono essere distinti in tre gruppi: modelli di tipo Machine Learning allenati su una rappresentazione documento-termine del testo (Support Vector Machines, Random Forest and Multinomial Näive Bayes), modelli di tipo Neural Network allenati sulle sequenze di testo (Long Short Term Memory) e modelli di tipo Machine Learning allenati su un insieme di caratteristiche linguistiche ricavate dal testo. Mentre i singoli modelli sono molto suscettibili all'overfitting, la classificazione ottenuta con il meccanismo di voto sembra essere più stabile ed ha mostrato una performance adeguata sul test set ufficiale del contest.*

Keywords
Conspiracy Theory, Content Moderation, Large Language Models, Computational Social Science

1. Introduction

With the proliferation of misinformation across multiple platforms and channels, detecting conspiracy theories and fake news has become a crucial task to ensure public safety and preserve democratic discourse. There have been efforts to develop Natural Language Processing (NLP) techniques to identify whether a given text contains fake news content [1] or not. The Automatic Conspiracy Theory Identification (ACTI) [2] task A in the EVALITA 2023 competition [3] aims to investigate different approaches to detect conspiracy theories from messages shared in platforms with lax moderation policies (like Telegram, 4chan, and Parler).

Different kinds of models are able to capture different aspects of a text and define suitable predictors for conspiracy messages.

Machine Learning (ML) models based on a matrix representation of terms in documents (document-term matrix) only "understand" which terms are used in a sentence. Therefore, they are not able to capture all features that can characterize a conspiratorial text.

Moreover, Artificial Neural Network (ANN) models does not only consider the presence of single words, but they are also able to estimate patterns and relationships among sequences of tokens in a sentence.

ML classifiers can also be trained on sets of linguistic features that can be predominantly present in fake content [4]. In literature, conspiratorial texts tends to display some particular linguistic properties [5] like both shortest and longest sentence, higher volume of punctuation marks, more frequent use of exclamation and question marks, a prominent use of adverbs [5, 6], predicative and attribute adjectives [7], capitalization and interjections [8, 9]. Emoji also seems to be more popular within fake news compared to true news [10], since they are used to increase persuasiveness of electronic communication [11].

Fake news corpus also tends to display also grammatical mistakes and inconsistencies, problems with sentence structure and an incorrect use of punctuation [5]. However, these latter features are more likely to be detected by a qualitative rather than quantitative analysis.

2. Related Work

Moderation of Fringe Communities. Recently, moderation of online fringe communities has become a pressing concern for mainstream platforms. Numerous works

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
selene.bianco@aizoongroup.com (S. Bianco);
daniela.salusso@unito.it (D. Salusso)
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper39.pdf

focused on the the efficacy of "deplatforming" these online communities [12, 13] as a way to limit the spreading of toxic ideas on mainstream platforms. However, when content moderation policies are applied users often migrate to alternative *fringe platforms*, sometimes created exclusively to host the banned community [14, 15]. Banning, in that context, would not only strengthen the infrastructure hosting these fringe platforms [12] but allow these communities to become more toxic elsewhere [16]. Indeed, such moderation policies do not avoid these fringe communities to grow and spread misinformation, for instance, about conspiratorial content. Therefore developing automated models to identify the spreading of such dangerous ideas is of the utmost importance.

Automated Methods for Identification of Toxic Content. Recent advances of natural language processing [17, 18, 19, 20] provides tools to address questions related to the identification of troublesome content online. Clearly traditional methods utilizing surface level features [21, 22] have opened the path to automatic detection of hate speech and fake news. Word embeddings [23] and recurrent neural networks [24, 25] significantly increased the prediction abilities of models to identify troublesome content. Finally, more recent systems based on transformers architectures [17, 18] have improved prediction accuracy among numerous tasks in different fields spanning from politics [26, 27], conflict prediction [28], and, of course, hate speech detection [29, 30, 31].

3. Description of the System

In order to evaluate different aspects of the text, different models have to been tested on a training-validation (70%-30%) split on the labeled training dataset. Since its dimension for task A is 1842 records, models shown a high tendency to overfit. Once the more performing models and their parameters were selected, the final models were trained again on the entire training set in order to predict labels on the test set (460 records).

In order to limit this problem and build a more stable system, we have defined a voting ensemble machine learning model combining the predictions from different models able to capture different features from the text. Therefore, the final ensemble classification represents the majority voting of the single predictions from each model.

3.1. ML models

Several ML models have been tested. The ones that performed better and have been selected as part of the voting approach were: Support Vector Machines (SVM) with radial basis function (rbf), Random Forest (RF) with maximum depth equal to 50 and Multinomial Näive Bayes (MNB).

In order to fit ML models, the training data has to be preprocessed and returned in the form of a document-term matrix. After cleaning the text from punctuation and stopwords, we tried to fit each model both on lemmatized and original tokens, once with only words and another with emoji too.

3.2. ANN model

The ANN model implemented is a Long Short-Term Memory (LSTM) with the structure as described in table 1.

3.3. ML model based on syntactical features

The latter model is trained on a set of syntactical features that have been extracted from the texts. The defined features were: the number of tokens and the percentages of punctuation, emoji, uppercase words, adjectives, adverbs, and interjections in the text. All the percentages had as denominators the sum of the number of tokens, punctuation signs, and emoji. The emoji were identified in the text with the emoji library. The Italian flag emoji was not recognized which this library and then identified in a second step. The Part Of Speech (POS) tagging used to count the number of adjectives, adverbs and interjections was performed with the Spacy's pipeline "it_core_news_sm".

We have also tried to identify the percentage of misspelled words with the phunspell library. However, the result was not entirely accurate and we decided to discard this feature for this reason.

We have first tested the significance of each feature with a probit logistic regression model and then used the significant ones as predictors in a RF model with maximum depth equal to 13.

4. Results and Discussion

The syntactical features used in the RF model are the ones that were significant in the logistic model (see table 2).

The logistic regression showed that a rise in the number of tokens and the proportion of punctuation and of uppercase words in the text is slightly increasing the probability of the text being of conspiracy type. These associations were in line with what was found in the literature [5]. On the contrary, the negative association of the proportion of emojis [10], adjectives [6], adverbs [5] and interjections [9] seems to be in contrast with what previously found for fake news in the English language.

Layer (type)	Output Shape	Param
embedding (Embedding)	(None, 645, 70)	869960
lstm (LSTM)	(None, 645, 100)	68400
pooling (GlobalMaxPooling1D)	(None, 100)	0
dense (Dense)	(None, 2)	202

Table 1
LSTM model summary.

feature	coef	std	z	P> $\|z\|$	[0.025, 0.975]
number of tokens	0.0036	0.000	7.700	0.000	0.003, 0.005
% punct	0.4955	0.240	2.068	0.039	0.026, 0.965
% uppercase words	1.8740	0.427	4.385	0.000	1.036, 2.712
% emoji	-3.2503	0.875	-3.716	0.000	-4.964, -1.536
% adjectives	-1.3083	0.460	-2.844	0.004	-2.210, -0.407
% adverbs	-2.1999	0.500	-4.400	0.000	-3.180, -1.220
% interjections	-11.7131	4.887	-2.397	0.017	-21.292, -2.134

Table 2
Significant syntactical features in the logistic regression.

Fitting the model on the text containing the emojis does not seem to give better results than fitting the model on the text alone. SVM and MNB models gave better results with the original tokens, while RF works better with lemmatized tokens. On the train-validation split, RF and SVM obtained higher precisions, whereas MNB higher recall.

The LSTM model was highly prone to overfitting starting from the second epoch. The f-1 scores of the tested models on the final test set are reported in table 3.

Limitations and further developments

Although multilingual frameworks for fake news detection are starting to be developed, a significant efficiency gap in modeling exists between English and languages other than English [32].

The model based on linguistic features should be improved by further detailing of the POS categories (adjectives, adverbs and interjections) in subcategories to better identify key predictors for fake news. A deeper analysis should be performed on emojis too, since they are known to play a role in the spread of misinformation by appealing to emotions [10]. The proportion of emojis in the text seems to be associated with an increased probability of the text not being of conspiracy type probably because emoji can express different kinds of feelings and should be divided into subtypes too.

Moreover, explainable AI (XAI) techniques could be further identified to underline specific tokens or other elements in the text influencing the model to classify a text as conspiracy or not.

The corpus provided for the ACTI task would have benefited from a data augmentation technique [33] or an integration from other corpora. However, we have not worked on this side, nor tested techniques of bootstrap or boosting. In addition, we have not tried to use more sophisticated models like the ones exploiting attention mechanism, both because of the size of the training set and because the fine-tuning of large language models would have required large GPU resources.

Ethics Statement

As we explore the potential benefits and limitations of using artificial intelligence (AI) to detect and categorize online conspiracy theories, it is important to consider the broader implications and risks involved in this area of research. While AI holds great promise for helping us better understand the spread of misinformation online and assist human moderators in identifying problematic content, it must not be deployed without considering its limitations in terms of accuracy of the results. Moreover, we need to remember that models trained on textual data can be potentially at risk of privacy leakage by adversarial attacks [34].

References

[1] R. Oshikawa, J. Qian, W. Y. Wang, A survey on natural language processing for fake news detection, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp.

Model	F1-score
RF	0.7538
MNB	0.7147
SVM rbf	0.7573
LSTM	0.7422
RF with linguistic features	0.7007
Majority voting	0.7918

Table 3
F1-score of the single models and of the ensemble on the test data.

6086–6093. URL: https://aclanthology.org/2020.lrec-1.747.

[2] G. Russo, N. Stoehr, M. H. Ribeiro, Acti at evalita 2023: Overview of the conspiracy theory identification task, 2023. arXiv:2307.06954.

[3] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[4] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, R. Mihalcea, Automatic detection of fake news, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 3391–3401. URL: https://aclanthology.org/C18-1287.

[5] R. Sousa-Silva, Fighting the fake: A forensic linguistic analysis to fake news detection, International Journal for the Semiotics of Law - Revue internationale de Sémiotique juridique 35 (2022) 2409 – 2433. doi:10.1145/322234.322243.

[6] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, Y. Choi, Truth of varying shades: Analyzing language in fake news and political fact-checking, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 2931–2937. URL: https://aclanthology.org/D17-1317. doi:10.18653/v1/D17-1317.

[7] J. Grieve, H. Woodfield, The Language of Fake News, Elements in Forensic Linguistics, Cambridge University Press, 2023. doi:10.1017/9781009349161.

[8] A. Aich, S. Bhattacharya, N. Parde, Demystifying neural fake news via linguistic feature-based interpretation, in: Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 6586–6599. URL: https://aclanthology.org/2022.coling-1.573.

[9] H. Allcott, M. Gentzkow, Social media and fake news in the 2016 election, Journal of Economic Perspectives 31 (2017) 211–36. URL: https://www.aeaweb.org/articles?id=10.1257/jep.31.2.211. doi:10.1257/jep.31.2.211.

[10] S. Suntwal, L. Brandimarte, S. A. Brown, Understanding the role of nonverbal tokens in the spread of online information, Proceedings of the 56th Hawaii International Conference on System Sciences (2023) 5484–5493. URL: https://hdl.handle.net/10125/103303.

[11] T. Maiberger, D. Schindler, N. Koschate-Fischer, Let's face it: When and how facial emojis increase the persuasiveness of electronic word of mouth, Journal of the Academy of Marketing Science (2023). doi:10.1007/s11747-023-00932-8.

[12] E. Zuckerman, C. Rajendra-Nicolucci, Deplatforming our way to the alt-tech ecosystem, Knight First Amendment Institute at Columbia University, January 11 (2021).

[13] G. Russo, L. Verginer, M. H. Ribeiro, G. Casiraghi, Spillover of antisocial behavior from fringe platforms: The unintended consequences of community banning, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 17, 2023, pp. 742–753.

[14] C. Dewey, Washington Post — These are the 5 subreddits Reddit banned under its game-changing anti-harassment policy, and why it banned them, https://wapo.st/3AO7pbl, 2016.

[15] G. Russo, M. Horta Ribeiro, G. Casiraghi, L. Verginer, Understanding online migration decisions following the banning of radical communities, in: Proceedings of the 15th ACM Web Science Conference 2023, WebSci '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 251–259. URL: https://doi.org/10.1145/3578503.3583608. doi:10.1145/3578503.3583608.

[16] M. Horta Ribeiro, S. Jhaver, S. Zannettou, J. Blackburn, G. Stringhini, E. De Cristofaro, R. West, Do platform migrations compromise content moder-

ation? evidence from r/the_donald and r/incels, Proceedings of the ACM on Human-Computer Interaction 5 (2021) 1–24.

[17] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

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

[19] T. Sun, A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, D. Mirza, E. Belding, K.-W. Chang, W. Y. Wang, Mitigating gender bias in natural language processing: Literature review, arXiv preprint arXiv:1906.08976 (2019).

[20] G. Russo, N. Hollenstein, C. C. Musat, C. Zhang, Control, generate, augment: A scalable framework for multi-attribute text generation, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 351–366. URL: https://aclanthology.org/2020.findings-emnlp.33. doi:10.18653/v1/2020.findings-emnlp.33.

[21] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on twitter, in: Proceedings of the NAACL student research workshop, 2016, pp. 88–93.

[22] H. Hosseinmardi, S. A. Mattson, R. Ibn Rafiq, R. Han, Q. Lv, S. Mishra, Analyzing labeled cyberbullying incidents on the instagram social network, in: Social Informatics: 7th International Conference, SocInfo 2015, Beijing, China, December 9-12, 2015, Proceedings 7, Springer, 2015, pp. 49–66.

[23] O. Levy, Y. Goldberg, Dependency-based word embeddings, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2014, pp. 302–308.

[24] B. Gambäck, U. K. Sikdar, Using convolutional neural networks to classify hate-speech, in: Proceedings of the first workshop on abusive language online, 2017, pp. 85–90.

[25] P. Badjatiya, S. Gupta, M. Gupta, V. Varma, Deep learning for hate speech detection in tweets, in: Proceedings of the 26th international conference on World Wide Web companion, 2017, pp. 759–760.

[26] G. Russo, C. Gote, L. Brandenberger, S. Schlosser, F. Schweitzer, Disentangling active and passive cosponsorship in the u.s. congress, ArXiv abs/2205.09674 (2022).

[27] J. Valvoda, T. Pimentel, N. Stoehr, R. Cotterell, S. Teufel, What about the precedent: An information-theoretic analysis of common law, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 2275–2288. URL: https://aclanthology.org/2021.naacl-main.181. doi:10.18653/v1/2021.naacl-main.181.

[28] M. Zhong, S. Dhuliawala, N. Stoehr, Extracting victim counts from text, arXiv preprint arXiv:2302.12367 (2023).

[29] P. Alonso, R. Saini, G. Kovács, Hate speech detection using transformer ensembles on the hasoc dataset, in: Speech and Computer: 22nd International Conference, SPECOM 2020, St. Petersburg, Russia, October 7–9, 2020, Proceedings, Springer, 2020, pp. 13–21.

[30] R. T. Mutanga, N. Naicker, O. O. Olugbara, Hate speech detection in twitter using transformer methods, International Journal of Advanced Computer Science and Applications 11 (2020).

[31] L. Stappen, F. Brunn, B. Schuller, Cross-lingual zero-and few-shot hate speech detection utilising frozen transformer language models and axel, arXiv preprint arXiv:2004.13850 (2020).

[32] R. Mohawesh, S. Maqsood, A. Qutaibah, Multilingual deep learning framework for fake news detection using capsule neural network, Journal of Intelligent Information Systems (2023) 1–17. doi:10.1007/s10844-023-00788-y.

[33] A. Keya, M. A. Wadud, M. Ph. D., M. Alatiyyah, M. A. Hamid, Augfake-bert: Handling imbalance through augmentation of fake news using bert to enhance the performance of fake news classification, Applied Sciences 12 (2022) 8398. URL: https://www.mdpi.com/2076-3417/12/17/8398. doi:10.3390/app12178398.

[34] L. Song, X. Yu, H.-T. Peng, K. Narasimhan, Universal adversarial attacks with natural triggers for text classification, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 3724–3733. URL: https://aclanthology.org/2021.naacl-main.291. doi:10.18653/v1/2021.naacl-main.291.

Cicognini at ACTI: Analysis of techniques for conspiracies individuation in Italian

Giacomo Cignoni[1], Alessandro Bucci[1]

[1]University of Pisa

Abstract

This report illustrates methods and results for solving SubtaskA (conspiracy detection) and SubtaskB (conspiracy topic classification) of EVALITA 2023 ACTI challenge. We employed different transformer-based models and an original method based on tf-idf. Results shows top performance scores over 80% for both subtasks.

Keywords

Conspiracy Theory, Content Moderation, Large Language Models, Computational Social Science

1. Introduction

We decided to cover the EVALITA 2023 challenge "Automatic Conspiracy Theory Identification" or ACTI for short [2]. This challenge is about classifying whenever an Italian message is conspiratorial or not and, if positive, what type of conspiracy is about. Therefore the challenge is subdivided is 2 subtasks:

- **Conspiratorial Content Classification**: the model must recognize if a telegram post is conspiratorial or not.
- **Conspiracy Category Classification**: the model must discriminate to which conspiracy theory a post belongs from a list of 4 possible conspiracy topics:
  1. Covid-Conspiracy
  2. Qanon-Conspiracy
  3. Flat Earth-Conspiracy
  4. Pro-Russia Conspiracy

2. Related works

Conspiratorial content has been raising on the internet over the past years such that some has define it as a "Golden Age of Conspiracy" [3]. Indeed mainstream platforms have tried to moderate the diffusion of online communities with the implementation of content moderation known as deplatforming. However, there have been a lot of discussion regarding the efficacy of such interventions [4, 5, 6].. Indeed, some identified the presence of spillover of toxic behaviour [7] and the the presence of a radicalization process after the application of content moderation [8]. Therefore, the need for automatic models that can detect the diffusion of troublesome (or more specifically) conspiratorial content has become crucial. Transformer based models have revolutionized modern natural language processing [9, 10, 11, 12]. Indeed, they are the current state of the art models in most NLP tasks spanning different fields from politics [13, 14], conflict prediction [15], and, of course, hate speech detection [16, 17, 18].[19]. In particular finetuning of BERT[20] based models for classification tasks such as sentiment analysis or topic detection has been widely studied and its effectiveness proved with multiple benchmarks [21]. The usage of machine learning techniques for detecting conspiracy theories has been studied mainly in regard to social media texts extracted in the English language, although also classification on different topic of the conspiracies has been considered [22, 23].

3. Datasets

The 2 provided datasets are a collection of labeled Italian Telegram's messages. Both datasets were relatively clean in regard to the text, so heavy preprocessing was not needed.

3.1. Subtask A dataset

More specifically for Subtask A, the training dataset is a .csv file containing:

- **id**: unique post identifier.
- **comment_text**: the text of the telegram's message.
- **conspiratorial**: a binary label that indicates if the message is conspiratorial or not.

The training dataset is composed by 1842 samples, of which 925 with a positive conspiratorial label and 917

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT [1]

g.cignoni3@studenti.unipi.it (G. Cignoni);
a.bucci12@studenti.unipi.it (A. Bucci)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper40.pdf

with a negative conspiratorial label. The hidden test set is composed by 460 samples instead.

3.2. Subtask B dataset

And for Subtask B, the training dataset is a .csv file containing:

- **id**: unique post identifier.
- **comment_text**: the text of the telegram's message.
- **conspiracy**: a label going from 0 to 3 indicating which conspiracy topic the message is about.

The training dataset is composed by 810 samples, with the following conspiracy label distribution: 435 Covid-Conspiracy, 242 Qanon-Conspiracy, 76 Flat Earth-Conspiracy, 57 Pro-Russia Conspiracy. The hidden test set is composed by 300 samples instead.

4. Models

Due to the nature of the tasks, we mainly decided to try different types of transformers based models for both subtasks, in order to capture the semantics and the general matter of the message itself. This is concatenated with a densely connected neural network in order to classify what the specific task is asking.

More specifically a Transformer as described in "Attention is all you need" [10], is composed of encoder-decoder structure composed by multiple modules stacked Nx times on top of each other like in Figure 1 where each module is mainly consisted of Multi Head Attentions and Feed Forward layers. In this architecture, the inputs and the outputs (target sentences) are embedded (the outputs need a right shift before usage) into an n-dimensional space because we cannot use the strings directly.

Here we present the selected transformer-based models for the tasks. Those were selected after a preliminary exploratory phase based on their performance on the validation set.

4.1. BERT-xxl

We used the *bert-base-italian-xxl-cased* model[24], which is an Italian pretrained BERT, an encoder-only transformer, variant developed by MDZ Digital Library team. It was pretrained using as source data a Wikipedia dump of various texts from the OPUS corpora collection with a size of 13 GB and more than 2 billion tokens. With the XXL variant, the corpus was extended with the Italian part of the OSCAR corpus, reaching a size of 81 GB and more than 13 billion tokens. This BERT-xxl model has 12 hidden layers, 12 attention heads and a hidden size of 768. We executed fine tuning on the transformer. Classification is executed on the first special output token [CLS] of the transformer

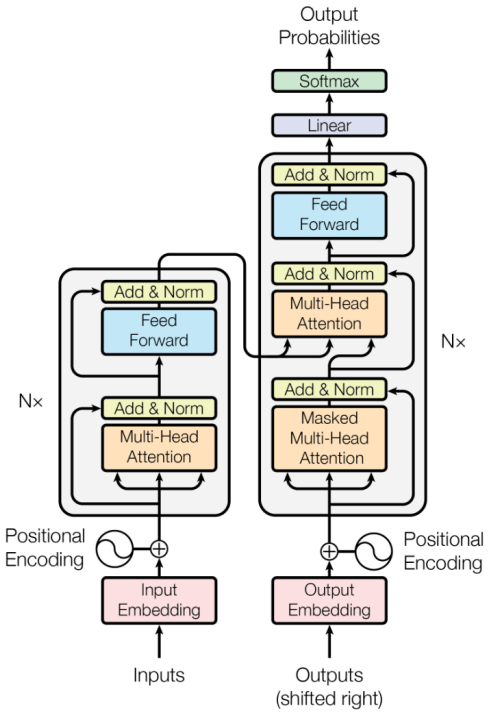


Figure 1: The transformer architecture.

4.2. XLM-RoBERTa

XLM-RoBERTa [25] is a multilingual version of RoBERTa, a transformers model pre-trained in a self-supervised fashion, similarly to BERT, but with a larger corpus and no next sentecnce prediction. *XLM-RoBERTa* was pretrained on 2.5TB of filtered CommonCrawl data containing 100 languages. Specifically, we used the *xlm-roberta-large* variant, which has 24 hidden layers, 16 attention heads and a hidden size of 1024. We executed fine tuning on the transformer. Classification is executed on the first special output token [CLS] of the transformer.

4.3. Llama

LLaMA is an autoregressive language model developed by Meta AI [26], based on a decoder only transformer architecture. We used the 7B variant, the smallest one, which has 7 billions parameters. It was pretrained on 1 trillion tokens from CCNet [67%], C4 [15%], GitHub

[4.5%], Wikipedia [4.5%], Books [4.5%], ArXiv [2.5%], Stack Exchange[2%] sources. The Wikipedia and Books sources are multilingual. Classification is executed on the last output token of the transformer.

We don't use fine tuning on this model due to its size, but only use it to generate sentence embeddings; trainining was only executed on the classification head.

4.4. Topic-specific tf-idf baseline

For Subtask B, considered its nature of topic classification and observing the presence of specific and unique words in each topic, we also developed an original heuristic baseline based on this assumptions. In short, it tries to retrieve the most specific keywords to each topic and extract their distribution in input texts. We recall that the definition of tf-idf for each word i in a set of documents $d \in D$ (in our case each document corresponds to each Telegram message in the dataset) is:

$$tf - idf_{i,j} = tf_{i,j} \times idf_i$$

, with $tf_{i,j} = \frac{n_{i,j}}{|d_j|}$ ($n_{i,j}$ being the number of occurrencies of word i in document j) and $idf_i = log_{10} \frac{|D|}{d:i \in D}$

This method makes use of *topic-specific tf-idf*, which is basically the normalized average tf-idf for each word in respect to the documents of each topic, then divided by the average tf-idf of the same word in the other topics. In mathematical terms, defining T as the set of topics, $avg_tfidf_{i,t}$ as the average tf-idf for word i and topic t, and $norm_avg_tfidf_{i,t}$ as the normalized $avg_tfidf_{i,t}$ in [0, 100] range, we have:

$$topic_specific_tfidf_{i,t} =$$

$$\frac{norm_avg_tfidf_{i,t}}{\sum_{t' \in T \setminus t} norm_avg_tfidf_{i,t'}}$$

This sccore is calculated only for the training set; for each topic t then we extract the top K $topic_specific_tfidf_{k,t}$ words and store them (K is an hyperparameter). Figure 2 shows the top 10 keywords for each topic with their respective score.

Finally, for each input text, we extract the distribution of the previously stored words, thus we obtain a $num_topics \times K$ distribution vector. This vector is then fed into a Random Forest (RF) model for the final classification. This model is trained with 6-fold Cross-Validation (CV) on the training set.

4.5. Preprocessing

For the transformer-based models, only light preprocessing was applied, only substituting break line characters with spaces and using each transformer tokenizer.

avg_tfidf_covid

	words	avg_tfidf_covid	avg_tfidf_qanon	avg_tfidf_russia	avg_tfidf_flat_earth
992	dos	23.939568	0.000000	0.0	0.0
316	avvers	15.754102	0.000000	0.0	0.0
1602	inocul	15.027083	0.000000	0.0	0.0
1967	mrna	14.798967	0.000000	0.0	0.0
1372	giovan	14.456688	0.000000	0.0	0.0
1039	efficac	13.687523	0.000000	0.0	0.0
1195	farmac	13.044677	0.000000	0.0	0.0
2055	obbligator	12.993569	0.000000	0.0	0.0
2964	sperimental	11.597360	0.000000	0.0	0.0
1832	mascherin	11.391805	0.000000	0.0	0.0

avg_tfidf_qanon

	words	avg_tfidf_covid	avg_tfidf_qanon	avg_tfidf_russia	avg_tfidf_flat_earth
3223	trump	0.020017	98.037534	0.0	0.0
958	dittatur	0.285784	48.449649	0.0	0.0
3822	elezion	0.000000	46.367399	0.0	0.0
1223	femmin	0.032511	39.571809	0.0	0.0
2737	sar	0.245433	30.175185	0.0	0.0
3821	elettoral	0.000000	29.785769	0.0	0.0
3871	fbi	0.000000	27.942181	0.0	0.0
3643	clinton	0.000000	27.854850	0.0	0.0
1252	finc	0.197767	27.317816	0.0	0.0
658	condivid	0.292505	27.184331	0.0	0.0

avg_tfidf_russia

	words	avg_tfidf_covid	avg_tfidf_qanon	avg_tfidf_russia	avg_tfidf_flat_earth
3239	ucrain	0.040250	0.338406	70.648510	0.0
2707	russ	0.025138	0.438085	68.342292	0.0
1992	nat	0.050197	0.191845	37.563651	0.0
4740	zelensky	0.000000	0.252306	30.421714	0.0
1910	milit	0.045818	0.232245	25.686416	0.0
2139	pac	0.051818	0.150653	24.721023	0.0
263	attacc	0.028973	0.412274	24.385013	0.0
4781	azov	0.000000	0.000000	20.561859	0.0
4197	nazist	0.000000	0.639016	18.741174	0.0
4990	polacc	0.000000	0.000000	17.956986	0.0

avg_tfidf_flat_earth

	words	avg_tfidf_covid	avg_tfidf_qanon	avg_tfidf_russia	avg_tfidf_flat_earth
3131	terr	0.051687	0.088669	0.062931	83.105710
4472	satell	0.000000	0.028737	0.000000	59.635025
2262	piatt	0.035781	0.000000	0.000000	58.062935
5160	lun	0.000000	0.000000	0.000000	57.843479
5173	orbit	0.000000	0.000000	0.000000	52.471631
3898	fot	0.000000	0.111569	0.185828	46.121322
4955	nas	0.000000	0.000000	0.048092	45.341877
2278	piu	0.051631	0.187217	0.050303	43.793724
1027	ecco	0.147374	0.230873	0.180290	43.415115
3933	glob	0.000000	0.045826	0.000000	42.738272

Figure 2: This figure reports the top 10 keywords for each of the topics (Covid, Qanon, pro-Russia and Flat-earth in descending order). Keywords are obtained using the *Topic-specific tf-idf* model. For each set of top words of each topic, also the score for the same words of the other topics is shown. It is easy to note that all top words have an high score for their respective topic, but very low ones for other topics.

Instead, for the *Topic-specific tf-idf* model, as the focus are topic specific relevant words, we apply stop word and short words (less than 3 characters) removal, number and punctuation elimination and stemming.

5. Implementation

We used the Python environment for developing the models, using mainly PyTorch, Scikit-Learn and Transformers libraries.

6. Experiments and results

We used an hold-out approach for both subtasks, reserving 20% of the training set for validation for hyperparameter tuning (split with labels ratio preservation). We experimented with retrain on validation found hyperparameters, but with worse results, so we decided to keep the model tested on the validation set as the final model for each configuration.

Model	Learning Rate	LR warmup ratio	Gamma	Head layers sizes
BERT-xxl	[1e-6, 2e-6, **3e-6**]	[**0.1**, 0.05]	-	[{128,2}, {256,2}, **{512,32,2}**]
RoBERTa-XLM	[6e-6, **8e-6**]	[**0.05**]	-	[{768,2}, **{1024,64,2}**]
Llama 7B	[**1e-5**, 5e-5, 1e-4, 5e-4]	-	[0.95, **0.99**]	[{4096,128,2}, {4096,512,32,2}, **{8192,1024,64,2}**]

Table 1
Subtask A hyperparameters, best found hyperparameters in bold.

Model	Learning Rate	LR warmup ratio	Gamma	Head layers sizes
BERT-xxl	[3e-6, **4e-6**, 5e-6]	[0.08, **0.04**]	-	[{256,4}, **{512,4}**, {1024,4}]
RoBERTa-XLM	[6e-6, **8e-6**]	[**0.05**]	-	[**{1536,4}**, {1536,64,4}]
Llama 7B	[**1e-4**, 3e-4, 5e-4, 8e-4, 1e-3]	-	[0.99, **0.999**]	[{4096,128,4}, **{4096,768,64,4}**, {8192,1024,128,4}]

Table 2
Subtask B hyperparameters, best found hyperparameters in bold.

For the Topic-specific tf-idf baseline, the validation set was used for finding the best K. After this we used a retrain strategy, in order to obtain a more general *topic_specific_tfidf* for words in each topic (RF classifier was also retrained with same CV hyperparameters)

The performance score of choice is macro-averaged F1 score, as it is the one also used to evaluate the challenge.

6.1. Hyperparameters grid search

Tables 1, 2 and 3 display the explored hyperparameters respectively for transformer-based models in SubtaskA, transformer-based models in SubtaskB and *Topic-specific tf-idf baseline* model. The final chosen hyperparameters are those which yield the best score on the validation set and are highlighted in bold.

6.2. Results

Tables 4 and 5 display the scores on both the internal validation set (the score used to choose the model with the best hyperparameters) and the hidden test set, respectively for SubtaskA and SubtaskB. Only macro-averaged F1 score is reported in the tables.

The whole hidden test set is split in public and private test sets by the competition rules; the final test score is obtained by weighted average (proportional each of the 2 test set sizes) of the public and private sets.

7. Discussions

For both tasks, the best performing models are the BERT-based ones, both the Italian BERT-xxl and XLM-RoBERTa, as their performance is close in F1 terms and are the top-2 performers in both subtasks. These results are a probable cause of the benefits of finetuning or of the encoder-only transformer architecture, versus the decoder and not finetuned Llama.

Among the relevant findings we include also that the transformer dimension does not influence the performance score; for example, although XLM-RoBERTa employs a larger architecture than BERT, they are comparable. The same reasoning applies when confronting with Llama 7B, which has at least an order of magnitude more parameters than the other transformers.

This indicates that the pre-training dataset (we recall that BERT-xxl is not multilingual and trained only in Italian) and the choice of finetuning have the greatest impact on performances.

In regard to the *Topic-specific tf-idf* model, it provides solid results in exchange for a lower computational cost, thanks to its strong assumptions of the importance of topic specific keywords in Subtask B.

It is also important to note that the samples correctly identified by *Topic-specific tf-idf* are not a strict subset of correctly identified samples by the BERT model, as the predictions on the test set have a divergence ratio of almost 25%, while there is a performance difference of less than 7%, meaning that a substantial set of "hard" (wrongly classified) samples for the transformer model are instead "easy" (correctly classified) for the *Topic-specific tf-idf* and vice versa. This implies that combining the 2 models in a meaningful way could result in a more robust model.

References

[1] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language process-

K	[10, 20, 30, 40, 50, 60, **70**, 80, 90, 100]
Random Forest max_depth	[5, 15, **None**]
Random Forest max_features	[**log2**, None]
Random Forest min_samples_leaf	[[1, **2**, 4]
Random Forest n_estimators	[[64, 128, **256**]

Table 3
Subtask B *Topic-specific tf-idf baseline* hyperparameters, best found hyperparameters in bold.

	Validation score	Test score
BERT-xxl	0.8184	**0.8257**
XLM-RoBERTa	0.7989	0.8203
Llama 7B	0.7954	0.8022

Table 4
Subtask A validation and test scores. Best test model is in bold.

	Validation score	Test score
BERT-xxl	0.8651	0.8265
XLM-RoBERTa	0.8776	**0.8532**
Llama 7B	0.8123	0.7389
Topic-specific tf-idf	0.7400	0.7520

Table 5
Subtask B validation and test scores. Best test model is in bold.

ing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[2] G. Russo, N. Stoehr, M. H. Ribeiro, Acti at evalita 2023: Overview of the conspiracy theory identification task, arXiv preprint arXiv:2307.06954 (2023).

[3] H. W. Hanley, D. Kumar, Z. Durumeric, A golden age: Conspiracy theories' relationship with misinformation outlets, news media, and the wider internet (preprint) (2023).

[4] E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, E. Gilbert, You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech, Proc. ACM Hum.-Comput. Interact. 1 (2017). URL: https://doi.org/10.1145/3134666. doi:10.1145/3134666.

[5] E. Chandrasekharan, S. Jhaver, A. Bruckman, E. Gilbert, Quarantined! examining the effects of a community-wide moderation intervention on reddit, ACM Transactions on Computer-Human Interaction (TOCHI) 29 (2022) 1–26.

[6] A. Trujillo, S. Cresci, Make reddit great again: assessing community effects of moderation interventions on r/the_donald, Proceedings of the ACM on Human-Computer Interaction 6 (2022) 1–28.

[7] G. Russo, L. Verginer, M. H. Ribeiro, G. Casiraghi, Spillover of antisocial behavior from fringe platforms: The unintended consequences of community banning, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 17, 2023, pp. 742–753.

[8] G. Russo, M. H. Ribeiro, G. Casiraghi, L. Verginer, Understanding online migration decisions following the banning of radical communities, arXiv preprint arXiv:2212.04765 (2022).

[9] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017. arXiv:1706.03762.

[11] T. Sun, A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, D. Mirza, E. Belding, K.-W. Chang, W. Y. Wang, Mitigating gender bias in natural language processing: Literature review, arXiv preprint arXiv:1906.08976 (2019).

[12] G. Russo, N. Hollenstein, C. C. Musat, C. Zhang, Control, generate, augment: A scalable framework for multi-attribute text generation, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 351–366. URL: https://aclanthology.org/2020.findings-emnlp.33. doi:10.18653/v1/2020.findings-emnlp.33.

[13] G. Russo, C. Gote, L. Brandenberger, S. Schlosser, F. Schweitzer, Disentangling active and passive cosponsorship in the u.s. congress, ArXiv abs/2205.09674 (2022).

[14] J. Valvoda, T. Pimentel, N. Stoehr, R. Cotterell, S. Teufel, What about the precedent: An information-theoretic analysis of common law, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 2275–2288. URL: https://aclanthology.org/2021.naacl-main.181. doi:10.18653/v1/2021.naacl-main.181.

[15] M. Zhong, S. Dhuliawala, N. Stoehr, Extract-

ing victim counts from text, arXiv preprint arXiv:2302.12367 (2023).
[16] P. Alonso, R. Saini, G. Kovács, Hate speech detection using transformer ensembles on the hasoc dataset, in: Speech and Computer: 22nd International Conference, SPECOM 2020, St. Petersburg, Russia, October 7–9, 2020, Proceedings, Springer, 2020, pp. 13–21.
[17] R. T. Mutanga, N. Naicker, O. O. Olugbara, Hate speech detection in twitter using transformer methods, International Journal of Advanced Computer Science and Applications 11 (2020).
[18] L. Stappen, F. Brunn, B. Schuller, Cross-lingual zero-and few-shot hate speech detection utilising frozen transformer language models and axel, arXiv preprint arXiv:2004.13850 (2020).
[19] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Huggingface's transformers: State-of-the-art natural language processing, 2020. arXiv:1910.03771.
[20] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.
[21] C. Sun, X. Qiu, Y. Xu, X. Huang, How to fine-tune bert for text classification?, 2020. arXiv:1905.05583.
[22] H. Guo, A. Ash, D. Chung, G. Friedland, Detecting conspiracy theories from tweets: Textual and structural approaches., in: MediaEval, 2020.
[23] W. Marcellino, T. C. Helmus, J. Kerrigan, H. Reininger, R. I. Karimov, R. A. Lawrence, Detecting Conspiracy Theories on Social Media: Improving Machine Learning to Detect and Understand Online Conspiracy Theories, RAND Corporation, Santa Monica, CA, 2021. doi:10.7249/RR-A676-1.
[24] DBMDZ, bert-base-italian-xxl-cased, 2020. URL: https://huggingface.co/dbmdz/bert-base-italian-xxl-cased.
[25] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, 2020. arXiv:1911.02116.
[26] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. arXiv:2302.13971.

Giobergia at Multi-Task Transformer Tuning for Joint Conspiracy Theory Detection and Classification

Flavio Giobergia

Department of Control and Computer Engineering, Politecnico di Torino, Turin, Italy

Abstract

Conspiracy theories, prevalent in contemporary society, often propagate misinformation and distrust, impacting public opinion and decision-making processes. In this paper, we present an automated approach to detect and classify conspiracy theories in Italian to address the Automatic Conspiracy Theory Identification (ACTI) task. Our methodology leverages a transformer-based architecture trained on a multi-task problem to tackle the challenging task of conspiracy theory identification. Through this multi-task learning framework, we aim to build a single model capable of addressing both the detection and the classification tasks simultaneously. We show that tackling both problems in a multi-task setting results in improved performance w.r.t. simple transformer-based solutions.

Keywords

transformers, conspiracy theory, deep learning, natural language processing

1. Introduction

Conspiracy theories have become a pervasive phenomenon, spreading through various communication channels, including social media platforms and online communities. These theories often involve the belief in secret plots or covert actions orchestrated by influential entities, which aim to manipulate events, control narratives, or conceal the truth. While some conspiracy theories may seem harmless or merely speculative, many have far-reaching consequences, potentially eroding trust in institutions, sowing discord among communities, and hindering informed decision-making processes: in recent years some of these conspiracies have been involved in the Capitol Hill attack (QAnon [1]) and hindered the efforts made toward the mitigation of the COVID-19 pandemic (e.g. resulting in lower vaccination and social distancing responses [2]), thus effectively jeopardizing lives.

The prevalence and impact of conspiracy theories necessitate effective methods for their detection and classification. Manual identification and analysis of conspiracy theories are time-consuming, resource-intensive, and subject to bias. Mainstream platforms (e.g. Reddit, Facebook) need to apply moderation policies at the community level. Given the limitations of the currently adopted approaches [3, 4], a more suitable methodology is required to address an efficient identification and classification of conspiracy theories that are constantly evolving. For these reasons, the Automatic Conspiracy Theory Identification (ACTI) [5], one of the tasks of EVALITA 2023 [6], is aimed at advancing the automation of these tasks for the Italian language. In particular, the task identifies two main goals: detecting whether a short piece of text is conspiratorial in nature or not and, if so, to which kind of conspiracy theories it conforms.

The detection of conspiracy theories in online contents is not a new one: other works have focused, for example, on the detection of conspiracies and misinformation related to COVID-19 [7], whereas the authors in [8, 9] propose building an automated pipeline for the detection of conspiracy theories and their diffusion by analyzing a network of actors and the interactions occurring within. To the best of our knowledge, ACTI is the first challenge on the detection of conspiracy theories with a specific focus on the Italian language. It should be pointed out, however, that the problem under study often spans across multiple languages, given the often English-centric origin of such theories. Various works have focused on the cross-lingual detection of adjacent topics, such as hate speech [10] and fake news [11], or sentiment analysis [12, 13], thus offering useful building blocks for future possible applications.

In this paper we propose leveraging Natural Language Processing (NLP) techniques to address the ACTI task. We employ a pre-trained transformer-based model fine-tuned to address both subtasks simultaneously. The source code for the proposed method is openly available on GitHub[1]

The remainder of this paper is organized as follows: Section 2 provides an overview of the subtasks and the data, Section 3 introduces the proposed method to address both subtasks and Section 4 presents the results obtained. Finally, Section 5 draws conclusions based on

[1] https://github.com/fgiobergia/EVALITA-ACTI-2023

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT

flavio.giobergia@polito.it (F. Giobergia)

0000-0001-8806-7979 (F. Giobergia)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper41.pdf

the results achieved.

2. Problem overview

The ACTI task is focused on the detection and classification of conspiracy theories based on short text posts. The data is collected from Telegram Channels and is entirely in Italian, with the exception of some short citations in English. In particular, two subtasks are proposed:

- *Subtask A*: the main goal is to detect whether a post is conspiratorial in nature or not, so the problem is framed as a binary classification one. A training set $\mathcal{D}_A$ containing a total of 1,841 posts is made available, with approximately 50% of the messages within being conspiratorial and 50% not. The metric used for the evaluation of this subtask is the macro F_1 score (i.e. the unweighted average F_1 score for the two classes). The unlabelled test set is instead comprised of a total of 460 posts.
- *Subtask B*: in this task, the posts provided can be classified as conforming to one of four conspiracy theories, namely Covid, QAnon, Flat Earth, Russia (more details on each conspiracy theory are provided in the task paper). Each post should be classified as belonging to either one of these categories[2] based on its contents. The training set $\mathcal{D}_B$ is comprised of 810 samples, with an approximate split of 50/30/10/10 among the Covid, QAnon, Flat Earth, Russia classes. The unlabelled test set contains 300 unlabelled samples. The macro F_1 score is used as the main evaluation metric for this problem.

Finally, we note that there are some overlaps in the posts contained in the two subtasks. While this is not, in general, a problem when the two training sets overlap (although it needs to be kept into account for a multi-task solution, to avoid data leakage during validation), we note that the test set of Subtask A has 164 posts in common with the training set and 66 with the test set of Subtask B, for a total of 230 posts that can be easily assumed to be conspiratorial in nature. In the spirit of a fair competition this information has obviously not been used in any way during the competition.

3. Methods

Let X be the set of all possible inputs (i.e. Telegram posts) and $\mathcal{C}$ be the set of possible conspiracies. We can formalize Subtask A as building a model $f_A : X \to [0, 1]$ that estimates the probabilities for a post to be conspiratorial in nature and Subtask B as a model $f_B : X \to [0, 1]^{|\mathcal{C}|}$ that estimates the distribution of probabilities across conspiracy classes (the following hold for all x: $f_B(x)_i \geq 0$ and $\sum_i f_B(x)_i = 1$). We note that, although the two functions produce different results, they work on the same inputs. We thus propose building an encoding function $e : X \to \mathbb{R}^d$ that projects the inputs into a shared latent space, and two head functions h_A, h_B such that $f_A = h_A \circ e$ and $f_B = h_B \circ e$. In other words, we aim to build robust shared representations by framing the problem as a multi-task one.

By introducing a common encoder, we can use two simple classification models for the heads, deferring the complexity of the entire model to $e(\cdot)$. In particular, we use a pre-trained transformer which is fine-tuned on the conspiracy theory detection and classification tasks simultaneously.

To solve this multi-task problem we adopt a loss function that evaluates the model based on the two separate targets. In particular, detecting whether a message is conspiratorial in nature is a binary problem that can be evaluated in terms of binary cross-entropy. For a given message x_i with binary conspiratorial label $y_i^{(A)}$, we define the conspiratorial loss function as:

$$\mathcal{L}^{(A)}(x_i, y_i^{(A)}) = -y_i^{(A)} log\sigma(f_A(x_i)) - (1 - y^{(A)}) log(1 - \sigma(f_A(x_i))) \quad (1)$$

Where $\sigma(\cdot)$ is the sigmoid function. By contrast, the loss function for the conspiracy theory classification is a multi-class classification problem. As such, we aim to minimize the cross-entropy between the predicted probability distribution and the ground truth value $y_i^{(B)}$, defined as follows:

$$\mathcal{L}^{(B)}(x_i, y_i^{(B)}) = \sum_j y_{ij}^{(B)} softmax(f_B(x_i))_j \quad (2)$$

We note that the datasets available for the two tasks are not the same, although some overlaps occur. Because of this, the two losses $\mathcal{L}^{(A)}$ and $\mathcal{L}^{(B)}$ cannot be computed for all points. In particular, points that are not conspiratorial in nature ($y^{(A)} = 0$) are not associated with any conspiracy theory, thus making the term $\mathcal{L}^{(B)}$ meaningless. Similarly, points that are conspiratorial in nature but only appear as a part of the dataset for Subtask A are not annotated with a ground truth label regarding the conspiracy theory to which they conform. Because of this, all points that belong exclusively to the dataset for Subtask A can only be evaluated in terms of $\mathcal{L}^{(A)}$. Instead, all points exclusively belonging to Subtask B are guaranteed to be conspiratorial in nature. Because of this,

[2]We note that some conspiracy theories are strongly related to one another (e.g. QAnon and Covid), so the identification of a single class may at times lead to ambiguous results.

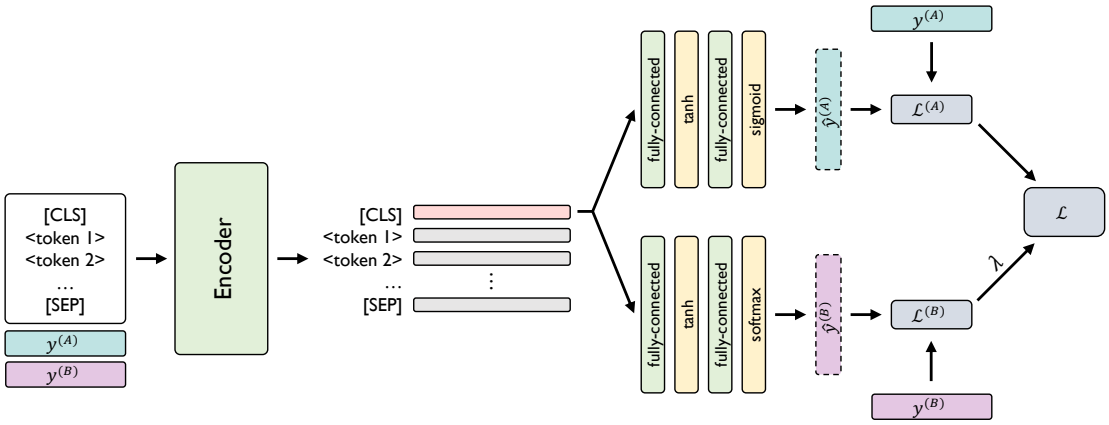


Figure 1: Architecture used for the proposed solution. In green are the trainable portions of the pipeline. The encoder used is transformer-based. The output vector for the [CLS] token is used as representative of the entire input sentence.

we infer for those points that $y^{(A)} = 1$. This consideration makes it reasonable to assume that the multi-task approach proposed should be particularly beneficial in terms of improvements on Subtask A.

The overall loss is obtained as a weighted sum of the above terms:

$$\mathcal{L} = \mathcal{L}^{(A)} + \lambda \mathbb{1}(x_i \in \mathcal{D}_B)\mathcal{L}^{(B)} \tag{3}$$

Where λ is used to balance the important that the two loss terms play in the overall predictions, and $\mathbb{1}(\cdot)$ is a selector that applies the second loss term only for terms where that portion can be applied meaningfully.

Figure 1 summarizes the architecture for the proposed methodology from inputs to the computation of the overall loss. Since we make use of transformer-based encoders, we note that we adopt the final hidden state corresponding to the [CLS] token as representative of the semantic contents of the entire message [14].

4. Results

The experimental section aims to evaluate the effectiveness of our proposed approach for detecting and classifying messages containing conspiracy theories. We present the main choices made in terms of encoder adopted, as well as results in terms of choice of hyperparameters.

4.1. Model evaluation

As already discussed, the models are evaluated on both subtasks by means of the macro F_1 score metric. This metric is particularly useful when evaluating models on unbalanced classes, as is for example the case with Subtask B.

We report some of the main results (i.e. encoders choice and multi-task vs single tasks comparison) in terms of performance on the final test set, whereas the other results on definition of λ in terms of performance on a validation set that has been obtained as a 20% hold out from the available dataset.

The test set made available for the competition is split intro a public and a private subsets (used for different parts of the competition itself). Both scores have been made available, for each submission, at the end of the challenge. For Subtask A, the private/public test is a 70/30 split, whereas for Subtask B the private/public test split is approximately 50/50.

For conciseness, we report an aggregated result that covers both private and public scores simultaneously with the aforementioned weights: more specifically, if $F^{(A)}_{1,public}$ is the score obtained on the public set for subtask A and $F^{(A)}_{1,private}$ is the score obtained on the private set, we report $F^{(A)}_1 = 0.3\, F^{(A)}_{1,public} + 0.7\, F^{(A)}_{1,private}$. Similarly, we report $F^{(B)}_1 = 0.5\, F^{(B)}_{1,public} + 0.7\, F^{(B)}_{1,private}$. We note that, despite assigning the same weights to the same partitions, the reported metrics are not the same as computing the F_1 scores on the entire test set – an operation that cannot be performed with the information at hand.

The overall metric used to evaluate the performance in the competition is a weighted average of the performance obtained on the two tasks, with weights 0.6 and 0.4 for Subtasks A and B respectively. Thus, we additionally report the overall score $F_1 = 0.6\, F^{(A)}_1 + 0.4\, F^{(B)}_1$.

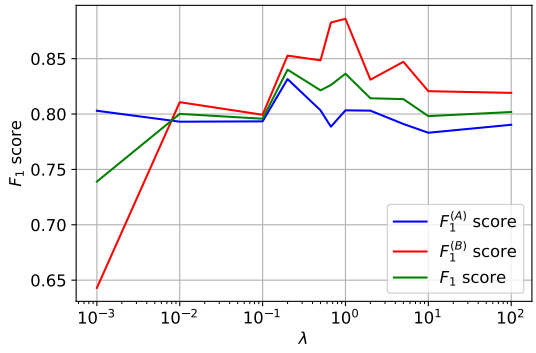


Figure 2: Performance of the model in terms of F_1 score (A, B, overall) as the parameter λ varies.

Model	$\mathbf{F_1^{(A)}}$	$\mathbf{F_1^{(B)}}$	$\mathbf{F_1}$
BERT-Italian-XXL-uncased	<u>0.8930</u>	<u>0.8475</u>	**0.8748**
BERT-Italian-XXL-cased	0.8686	0.8363	0.8556
BERT-Italian-base-uncased	**0.8953**	0.8082	0.8605
BERT-Italian-base-cased	0.8867	0.7920	0.8488
BART-IT-WITS	0.8494	0.8431	0.8469
BART-IT-IlPost	0.8504	0.8373	0.8452
BART-IT-FanPage	0.8327	**0.8490**	0.8393
BART-IT	0.8513	0.8322	0.8437
ELECTRA-XXL-discriminator	0.8886	0.8260	<u>0.8635</u>
ELECTRA-XXL-generator	0.8622	0.7693	0.8251

Table 1
Macro F_1 scores for the various encoding models used. Performance measured on subtasks A and B separately, as well as with the overall score (computed as $F_1 = 0.6\ F_1^{(A)} + 0.4\ F_1^{(B)}$). In **bold** are the best performing models for each metric. <u>Underlined</u> are the second best models.

4.2. Hyperparameters tuning

The main hyperparameters to be configured for the proposed pipeline is the λ coefficient. Other parameters that are generally important, but not specific to this precise context (e.g., number of training epochs, learning rate, layer sizes, optimizer) will not be covered in detail and can be found as a part of the provided source code.

The choice of a valid value for λ is quite specific to the conspiracy theory detection problem, as it represents the trade-off coefficient between the capability of detecting conspiratorial contents and being able to correctly assign them. Figure 2 shows how the performance of the model (based on BERT-Italian-XXL-uncased – as explained in the next subsection) varies as λ increases. Although the best value in terms of overall F_1 score is observed for $\lambda = 0.2$, we identify $\lambda = 1$ as being a more robust choice, considering the overall optimal behavior in the interval centered around that value.

4.3. Encoder choice

In recent years a wide variety of transformer-based models have been introduced for various languages, including Italian. An *a priori* choice regarding the most suitable model is not trivial to make. Therefore, we ran a benchmark study to assess how well various models behave on both tasks. In particular, we compare results obtained for the following models:

- **Italian ELECTRA** [15], an encoder-only architecture based on the ELECTRA model [16]; a method trained on detecting corruptions of the input introduced using a generator network. We test both the generator and discriminator versions of Italian ELECTRA.
- **BART-IT** [17], a sequence-to-sequence model based on the BART [18] architecture that is specifically tailored to the Italian language. BART-IT has been shown to outperform other state-of-the-art architectures on various tasks. We note that, on top of the original BART-IT model, three additional versions have been fine-tuned on abstractive summarization tasks on various datasets: FanPage, IlPost [19] and WITS (Wikipedia for Italian Text Summarization) [20]. Since these three data sources have rather different scopes and styles, we assess the quality of the various fine-tuned versions.
- **Italian BERT** [21], a BERT-based model [14] trained on a recent Wikipedia dump as well as data from the OPUS corpora[3] collection. We use both a cased and uncased version as well as a base and an XXL ones.

All models have been fine-tuned for a total of 11 epochs, considering that no meaningful improvement in performance on the validation set has been observed after this point. To reduce the computational cost, we only attempted a value for $\lambda = 1$ during the training of all models.

Table 1 presents the results obtained on the 10 models that have been tested. The best-performing model, in most cases, is BERT-Italian-XXL-uncased. The rest of the experiments presented will be performed using this encoder only, as a way to reduce the computational cost required.

4.4. Multi-task effect

Table 2 shows the results that are achieved by the multi-task model in contrast to the ones that can be obtained

[3] https://opus.nlpl.eu/

Task	$\mathbf{F_1^{(A)}}$	$\mathbf{F_1^{(B)}}$	$\mathbf{F_1}$
Multi-task	**0.8930**	0.8475	**0.8748**
A only	0.8370	-	0.8491*
B only	-	**0.8672**	0.8491*

Table 2
Results in terms of F_1 score when addressing the problem with a multi-task approach, or as separate subtasks. In **bold** are the best performing models for each metric. (* The overall results for "A only" and "B only" are obtained by merging the results obtained on the two subtasks separately.)

Submission	$\mathbf{F^{(A)}_{1,private}}$	$\mathbf{F^{(A)}_{1,public}}$	$\mathbf{F^{(B)}_{1,private}}$	$\mathbf{F^{(B)}_{1,public}}$
Submitted	0.8371	0.8389	0.8470	0.8360
Proposed	0.8907	0.8984	0.8805	0.8145

Table 3
Results in terms of macro F_1 score on the private and public sets, for both subtasks. Both the performance that have been submitted at the end of the challenge ("Submitted") and the best ones achieved ("Proposed") are shown.

by training the same model on only one of the tasks at a time. As expected, we observe a significant improvement in performance for Subtask A, whereas the performance of Subtask B does not benefit from the introduction of a multi-task approach. This can be explained in terms of benefits that are introduced by the multi-task loss: while points belonging to $\mathcal{D}_B$ could all be labelled as being conspiratorial (thus enhancing the dataset available for subtask A), points in $\mathcal{D}_A$ are not beneficial to the improvement of Subtask B (as they are either non-conspiratorial, or the conspiracy theory to which they conform is unknown).

4.5. Competition performance

The pipeline presented in this work is similar to the one used to take part to the competition, with some minor changes mainly regarding the adoption of BERT-Italian-XXL-uncased instead of BART-IT, as well as *not* adopting a different convolutional model in parallel. For the sake of completeness, the private scores obtained during the challenge are reported in Table 3, along with the best ones obtained with the proposed method.

5. Conclusions

In this work we presented a transformer-based multi-task approach to addressing a joint detection and classification problem. We have shown the importance of choosing the most suitable encoding model, as well as the benefits of adopting a multi-task approach. We highlighted how the current framing of the multi-task problem is only beneficial to one of the subtasks. As a future work, we aim to reframe the problem through a semi-supervised approach that allows for the pseudo-labelling of all points, thus aiming to improve the performance uniformly across subtasks.

Acknowledgments

This study was carried out within the FAIR - Future Artificial Intelligence Research and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013), with partial support from SmartData@PoliTO center on Big Data and Data Science. This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

References

[1] S. Moskalenko, C. McCauley, Qanon, Perspectives on Terrorism 15 (2021) 142–146.

[2] K. Bierwiaczonek, A. B. Gundersen, J. R. Kunst, The role of conspiracy beliefs for covid-19 health responses: A meta-analysis, Current Opinion in Psychology (2022) 101346.

[3] G. Russo, M. Horta Ribeiro, G. Casiraghi, L. Verginer, Understanding online migration decisions following the banning of radical communities, in: Proceedings of the 15th ACM Web Science Conference 2023, WebSci '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 251–259. URL: https://doi.org/10.1145/3578503.3583608. doi:10.1145/3578503.3583608.

[4] G. Russo, L. Verginer, M. H. Ribeiro, G. Casiraghi, Spillover of antisocial behavior from fringe platforms: The unintended consequences of community banning, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 17, 2023, pp. 742–753.

[5] G. Russo, N. Stoehr, M. H. Ribeiro, Acti at evalita 2023: Overview of the conspiracy theory identification task, arXiv preprint arXiv:2307.06954 (2023).

[6] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[7] S. Shahsavari, P. Holur, T. Wang, T. R. Tangherlini, V. Roychowdhury, Conspiracy in the time of corona: automatic detection of emerging covid-19 conspiracy theories in social media and the news, Journal of computational social science 3 (2020) 279–317.

[8] T. R. Tangherlini, S. Shahsavari, B. Shahbazi, E. Ebrahimzadeh, V. Roychowdhury, An automated pipeline for the discovery of conspiracy and conspiracy theory narrative frameworks: Bridgegate, pizzagate and storytelling on the web, PloS one 15 (2020) e0233879.

[9] G. Russo, C. Gote, L. Brandenberger, S. Schlosser, F. Schweitzer, Helping a friend or supporting a cause? disentangling active and passive cosponsorship in the U.S. congress, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2952–2969. URL: https://aclanthology.org/2023.acl-long.166.

[10] E. W. Pamungkas, V. Basile, V. Patti, A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection, Information Processing & Management 58 (2021) 102544.

[11] J. Köhler, G. K. Shahi, J. M. Struß, M. Wiegand, M. Siegel, T. Mandl, M. Schütz, Overview of the clef-2022 checkthat! lab task 3 on fake news detection, Working Notes of CLEF (2022).

[12] F. Giobergia, L. Cagliero, P. Garza, E. Baralis, Cross-lingual propagation of sentiment information based on bilingual vector space alignment., in: EDBT/ICDT Workshops, 2020, pp. 8–10.

[13] G. Russo, N. Hollenstein, C. C. Musat, C. Zhang, Control, generate, augment: A scalable framework for multi-attribute text generation, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 351–366. URL: https://aclanthology.org/2020.findings-emnlp.33. doi:10.18653/v1/2020.findings-emnlp.33.

[14] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[15] R. Guarasci, A. Minutolo, E. Damiano, G. De Pietro, H. Fujita, M. Esposito, Electra for neural coreference resolution in italian, IEEE Access 9 (2021) 115643–115654.

[16] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, Electra: Pre-training text encoders as discriminators rather than generators, arXiv preprint arXiv:2003.10555 (2020).

[17] M. La Quatra, L. Cagliero, Bart-it: An efficient sequence-to-sequence model for italian text summarization, Future Internet 15 (2023). URL: https://www.mdpi.com/1999-5903/15/1/15. doi:10.3390/fi15010015.

[18] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, arXiv preprint arXiv:1910.13461 (2019).

[19] N. Landro, I. Gallo, R. La Grassa, E. Federici, Two new datasets for italian-language abstractive text summarization, Information 13 (2022) 228.

[20] S. Casola, A. Lavelli, Wits: Wikipedia for italian text summarization., in: CLiC-it, 2021.

[21] S. Schweter, Italian bert and electra models, 2020. URL: https://doi.org/10.5281/zenodo.4263142. doi:10.5281/zenodo.4263142.

NERMuD at EVALITA 2023: Overview of the Named-Entities Recognition on Multi-Domain Documents Task

Alessio Palmero Aprosio[1,*,†], Teresa Paccosi[1,2,†]

[1]Fondazione Bruno Kessler, Via Sommarive 18, I-38121 Trento, Italy
[2]Dipartimento di Psicologia e Scienze Cognitive, Università di Trento, Corso Bettini 84, I-38068 Rovereto (TN), Italy

Abstract

In this paper, we describe NERMuD, a Named-Entities Recognition (NER) shared task presented at the EVALITA 2023 evaluation campaign. NERMuD is organized into two different sub-tasks: a domain-agnostic classification and a domain-specific one. We display the evaluation of the system presented by the only task participant, ExtremITA. ExtremITA proposes a unified approach for all the tasks of EVALITA 2023, and it addresses in our case only the domain-agnostic sub-task. We present an updated version of KIND, the dataset distributed for the training of the system. We then provide the baselines proposed, the results of the evaluation, and a brief discussion.

Keywords

Shared Task, Named-Entity Recognition, NERMuD 2023, EVALITA 2023

1. Introduction and Motivation

Named-entity recognition (NER) is one of the most common and important task in the field of Natural Language Processing (NLP). It involves identifying and classifying mentions of entities in texts and it is widely used in applications such as text understanding [1], information retrieval [2], knowledge base construction [3], and the protection of personal data [4]. These entities can belong to a set of predefined categories, with people, locations, and organizations being the most common ones.

Manually annotated data play a crucial role in training and evaluating NER systems, similar to other NLP tasks. Systems trained on datasets from specific domains often do not perform well when applied to different types of texts [5].

NER has been addressed in almost all languages, indicating a significant interest in the topic [6]. It is an important task in its own right, as it can be used to process large archival collections. While NER is considered a solved task, some studies have shown that there is always room for improvement depending on factors such as labels, languages, and topics [7]. It is worth noting that, despite the great number of studies on this topic, datasets and tasks for NER often focus on news and, more recently, social media, as seen in initiatives like I-CAB ([8]), NEEL-IT 2016 [9] and NER 2011 [10].

The rest of this article is structured as follows. Section 2 describes the task, and Section 3 gives an overview of the dataset provided. In Section 4 we portray the baseline and the evaluation metric, while in Section 5 we describe the work of the participant ExtremITA. In the end, Section 6 contains a brief discussion, while in Section 7 we draw some conclusions.

2. Task description

In this Section, we describe NERMuD, a task presented at EVALITA 2023 [11] that involves the extraction and classification of named entities – including persons, organizations, and locations – from documents in various domains.

NERMuD 2023 includes two different sub-tasks:

- **Domain-agnostic classification (DAC).** Participants are required to select and classify entities into three categories (person, organization, location) from different types of texts (news, fiction, political speeches) using a single general model.
- **Domain-specific classification (DSC).** Participants are required to make use of a different model for each of the above types, trying to increase the accuracy of every considered type.

The two classification tasks can be addressed in several ways. For example, using deep learning techniques or by adding external data such as gazetteers.

Participants are required to submit their runs and to provide a technical report that should include a brief description of their approach, focusing on the adopted algorithms, models, and resources, a summary of their experiments, and an analysis of the obtained results.

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
*Corresponding author.
†These authors contributed equally.
aprosio@fbk.eu (A. Palmero Aprosio); tpaccosi@fbk.eu (T. Paccosi)
0000-0002-1484-0882 (A. Palmero Aprosio); 0009-0009-2348-7556 (T. Paccosi)
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper42.pdf

Each participant can submit up to 3 runs for each subtask.

The runs should be contained in a TSV file with fields delimited by a tab and it should follow the same format of the training dataset. No missing data are allowed: a label should be predicted for each token in the test set.

3. Available dataset

The corpus that can be used for training is the Kessler Italian Named-entities Dataset (KIND) [12], presented in 2021 at the Language Resources and Evaluation Conference (LREC). KIND is available and freely downloadable on Github.[1]

The original dataset comprises over one million tokens and includes annotations for three entity classes: person, location, and organization. The majority of the dataset, approximately 600K tokens, features manual gold annotations across three distinct domains: news, literature, and political discourses. This specific subset can be used as the training data for the NERMuD 2023 task, which focuses on Named Entity Recognition and Multi-domain Classification.

All the texts used for the annotation are publicly available, under a license that allows both research and commercial use. In particular, the texts used for the NERMuD task come from:

- Wikinews (WN) as a source providing news texts from the last few decades;
- Some Italian fiction books (FIC) in the public domain, freely accessible for use;
- Writings and speeches from the Italian politician Alcide De Gasperi (ADG), a collection of texts including the works and speeches of Alcide De Gasperi, the Italian politician.

Since the dataset is already publicly released and available, a new set of data has been annotated and shared using the same guidelines (available on the KIND repository on Github).

The dataset has been collected in full compliance with ethical standards, ensuring that it aligns with the terms of use of the sources and that respects the intellectual property and privacy rights of the original authors of the texts.

Table 1 displays an overview of the dataset.

In the next subsections, we provide a quick description of the domains included in the dataset. For more information about the creation of the dataset, the text processing, and the annotation guidelines please refer to [12].

[1] https://github.com/dhfbk/KIND

3.1. Wikinews (WN)

Wikinews is a multi-language free-content project of collaborative journalism. The Italian chapter contains more than 11,000 news articles,[2] released under the Creative Commons Attribution 2.5 License.[3]

In building the dataset, we randomly choose 1,198 articles evenly distributed in the last 20 years, for a total of 364,816 tokens.

3.2. Literature (FIC)

For the annotation of fiction literature, we have included 86 book chapters from a collection of 11 publicly available Italian-authored books. This annotated dataset comprises a total of 219,638 tokens. While the majority of the selected books are novels, we have also included a mix of epistles and biographies. The plain texts come from the Liber Liber website.[4]

In particular, we select: *Il giorno delle Mésules* (Ettore Castiglioni, 1993, 12,853 tokens), *L'amante di Cesare* (Augusto De Angelis, 1936, 13,464 tokens), *Canne al vento* (Grazia Deledda, 1913, 13,945 tokens), *1861-1911 - Cinquant'anni di vita nazionale ricordati ai fanciulli* (Guido Fabiani, 1911, 10,801 tokens), *Lettere dal carcere* (Antonio Gramsci, 1947, 10,655), *Anarchismo e democrazia* (Errico Malatesta, 1974, 11,557 tokens), *L'amore negato* (Maria Messina, 1928, 31,115 tokens), *La luna e i falò* (Cesare Pavese, 1950, 10,705 tokens), *La coscienza di Zeno* (Italo Svevo, 1923, 56,364 tokens), *Le cose più grandi di lui* (Luciano Zuccoli, 1922, 20,989 tokens), *L'occhio del lago* (Tullio Giordana, 1899, 27,190 tokens).

We prioritized selecting texts in the public domain that are as recent as possible (considering that, under the current legislation, copyright expires 70 years after the death of the author). This choice was made to ensure that the model trained on this data would be well-suited for applying to novels written in recent years. By focusing on more contemporary texts, the language used in these novels is expected to be more similar to the language used in present-day novels. Additionally, for the test data, we specifically chose works by the author Tullio Giordana. His works are then not included in the train or the dev sets, to not have a model possibly biased in terms of style.

3.3. Alcide De Gasperi's Writings (ADG)

Finally, we annotate 173 documents (164,537 tokens) from the corpus described in [13], spanning 50 years of European history. The corpus is composed of a comprehensive collection of Alcide De Gasperi's public documents, 2,762

[2] https://it.wikinews.org/wiki/Speciale:Statistiche
[3] https://creativecommons.org/licenses/by/2.5/
[4] https://www.liberliber.it/

Table 1
Overview of the dataset

Dataset	Tokens	Train				Dev				Test			
		Total	LOC	PER	ORG	Total	LOC	PER	ORG	Total	LOC	PER	ORG
WikiNews	364,816	249,077	6,862	8,928	7,593	59,220	1,711	1,802	1,823	56,519	1,310	2,322	1,992
Fiction	219,638	170,942	733	3,439	182	21,506	463	636	284	27,190	37	443	1
De Gasperi	164,537	123,504	1,046	1,129	2,396	27,128	274	253	533	13,905	107	226	326
Total	748,991	543,523	8,641	13,496	10,171	107,854	2,448	2,691	2,640	97,614	1,454	2,991	2,319

in total, written or transcribed between 1901 and 1954, and it is available for consultation on the *Alcide Digitale* website.[5]

4. Baseline and Evaluation

During the definition of the task, we proposed two baselines: an old-style Conditional Random Field [14], and a plain BERT [15] implementation. These options represent the most effective algorithms that can be implemented without the use of GPUs, as well as the simplest algorithms that can be performed using transformers. Both implementations of the baselines can be found on Github.[6]

The CRF model is based on the classifier available in scikit-learn out-of-the-box. In addition to standard features extracted from the text, including vector information from fastText models [16], we also used a set of gazetteers (list of persons, organizations and locations) collected from the Italian Wikipedia using some of the classes contained in DBpedia [17]: Person, Organization, and Place, respectively.

The BERT NER classification model is inspired by the blog post of Tobias Sterbak,[7] using BertForTokenClassification[8] from Hugging Face.

Final results will be calculated in terms of macro-average F_1. The evaluation script is released in the KIND official Github project.[9]

5. Participants

The task has only one participant, the "ExtremITA" group [18], who participated in all the tasks presented at EVALITA 2023 with two unified multi-task learning approaches.

The purpose of ExtremITA is to investigate how the adoption of a Large Language Model can be taken to its extreme consequences by proposing a single model capable of tackling a wide array of heterogeneous tasks (among them, NERMuD).

The authors tested two different models:

extremIT5 - An Encoder-Decoder model based on IT5 [19] consisting of approximately 110 million parameters. This model is trained by concatenating the name of the task and the input sentence/paragraph in the input texts, each representing an example from a generic EVALITA task. Its purpose is to generate a piece of text that solves the target task. For NERMuD, in particular, the list of expected Named Entities is reported as a sequence of text spans, each associated with the corresponding entity type (in the form [⟨ENTITY_TYPE⟩] ⟨TEXT_SPAN_THAT_EVOKES_ENTITY⟩).

extremITLLaMA - An instruction-tuned Decoder-only model, built upon the LLaMA foundational models [20], with a total of 7k million parameters. The initial model was trained using the LoRA technique [21] on Italian translations of Alpaca [22] instruction data. The adapters are then merged into the original model. A final fine-tuning phase using LLaMA is then performed. For each example from EVALITA, an input text is paired with a manually crafted question that simulates an instruction to be solved, representing the specific task. The natural language instruction used in NERMuD is *"Scrivi le menzioni di entità nel testo, indicandone il tipo: [PER] (persona), [LOC] (luogo), [ORG] (organizzazione)."*[10].

In both cases, NERMuD was transformed into a sequence-to-sequence task from its original token classification format.

6. Discussion

Table 2 displays all the results on the different sets available for test. It is worth noting that the ExtremITA's unified approach is very similar to the baseline.

[5] https://alcidedigitale.fbk.eu/
[6] https://github.com/dhfbk/bert-ner
[7] https://bit.ly/ner-bert
[8] https://bit.ly/BertForTokenClassification
[9] https://github.com/dhfbk/KIND

[10] Write the entities' mentions in the text, indicating their type: [PER] (person), [LOC] (location), and [ORG] (organization)

Table 2
Evaluation of the two baselines and of the two ExtremITA runs.

Algorithm	Test	PER			LOC			ORG			Micro			Macro		
		P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Baseline (CRF)	ADG	0.93	0.85	0.89	0.84	0.88	0.86	0.79	0.57	0.66	0.85	0.72	0.78	0.85	0.77	0.80
Baseline (CRF)	WN	0.95	0.95	0.95	0.82	0.76	0.79	0.83	0.77	0.80	0.88	0.84	0.86	0.87	0.83	0.85
Baseline (CRF)	FIC	0.92	0.69	0.79	0.19	0.54	0.28	–	–	–	0.74	0.67	0.70	0.37	0.41	0.35
Baseline (CRF)	ALL	0.95	0.90	0.92	0.78	0.77	0.77	0.83	0.74	0.78	0.87	0.82	0.84	0.85	0.80	0.83
Baseline (BERT)	ADG	0.93	0.92	0.93	0.80	0.90	0.85	0.83	0.67	0.74	0.86	0.79	0.83	0.85	0.83	0.84
Baseline (BERT)	WN	0.96	0.98	0.97	0.89	0.87	0.88	0.87	0.88	0.87	0.91	0.92	0.91	0.91	0.91	0.91
Baseline (BERT)	FIC	0.95	0.82	0.88	0.51	0.76	0.61	–	–	–	0.90	0.81	0.85	0.49	0.52	0.50
Baseline (BERT)	ALL	0.95	0.95	0.95	0.87	0.87	0.87	0.87	0.85	0.86	0.91	0.90	0.90	0.90	0.89	0.89
ExtremIT5	ADG	0.89	0.87	0.88	0.78	0.85	0.81	0.75	0.58	0.65	0.81	0.72	0.76	0.80	0.77	0.78
ExtremIT5	WN	0.96	0.83	0.89	0.89	0.79	0.84	0.80	0.77	0.79	0.88	0.80	0.84	0.88	0.80	0.84
ExtremIT5	FIC	0.96	0.87	0.91	0.62	0.76	0.68	–	–	–	0.92	0.86	0.89	0.53	0.54	0.53
ExtremIT5	ALL	0.95	0.84	0.89	0.87	0.80	0.83	0.80	0.74	0.77	0.88	0.79	0.84	0.87	0.79	0.83
Extr.ITLLaMA	ADG	0.94	0.87	0.91	0.77	0.87	0.82	0.82	0.56	0.66	0.85	0.72	0.78	0.84	0.77	0.79
Extr.ITLLaMA	WN	0.96	0.96	0.96	0.91	0.84	0.87	0.86	0.83	0.84	0.91	0.88	0.90	0.91	0.87	0.89
Extr.ITLLaMA	FIC	0.97	0.82	0.89	0.81	0.81	0.81	–	–	–	0.96	0.82	0.88	0.59	0.54	0.57
Extr.ITLLaMA	ALL	0.96	0.93	0.95	0.89	0.84	0.86	0.86	0.79	0.82	0.91	0.86	0.89	0.90	0.85	0.88

The evaluation of ORG entities for the fiction domain is missing, as none of the classifiers were able to correctly identify the only ORG entity present in the test set (the work "Borsa" in the sentence "Ha avuto disgrazie alla Borsa"). Overall, the BERT baseline outperforms ExtremITA in most runs, with the exception of LOC extraction in fictional texts, where ExtremITLLaMA performs better. This difference in performance can likely be attributed to the textual data used to train the models.

In general, it is possible to notice that the best ExtremITA run overcomes almost always the classification in terms of precision.

7. Conclusions

In this paper we described the first evaluation task for multi-domain named-entity recognition in Italian texts. The task evaluated the performance of participant systems in terms of extracting entities that refers to persons, organizations, and location. The texts used for the tasks cover three different domains: news, political speeches, fiction.

Unfortunately, the task attracted only one participant, ExtremITA, who however presented an interesting and very innovative multi-task approach, probably the first one dealing with so many different tasks in Italian. Although in general the results of ExtremITA do not overcome the two strong baselines proposed (CRF w/ gazetteers, and BERT), the difference in terms of F_1 is very small, demonstrating a promising future for that kind of approaches.

As an outcome of the task, a new version of the KIND dataset is released, increasing its size with respect to the previous version.

References

[1] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, Q. Liu, ERNIE: Enhanced language representation with informative entities, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 1441–1451. URL: https://aclanthology.org/P19-1139. doi:10.18653/v1/P19-1139.

[2] J. Guo, G. Xu, X. Cheng, H. Li, Named entity recognition in query, in: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '09, Association for Computing Machinery, New York, NY, USA, 2009, p. 267–274. URL: https://doi.org/10.1145/1571941.1571989. doi:10.1145/1571941.1571989.

[3] O. Etzioni, M. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, A. Yates, Unsupervised named-entity extraction from the web: An experimental study, Artificial Intelligence 165 (2005) 91–134. URL: https://www.sciencedirect.com/science/article/pii/S0004370205000366. doi:https://doi.org/10.1016/j.artint.2005.03.001.

[4] T. Paccosi, A. Palmero Aprosio, Redit: A tool and

dataset for extraction of personal data in documents of the public administration domain, in: Proceedings of CLiC-it 2021 Italian Conference on Computational Linguistics, 2022.

[5] T. Poibeau, L. Kosseim, Proper Name Extraction from Non-Journalistic Texts, Brill, Leiden, The Netherlands, 2001, pp. 144 – 157. URL: https://brill.com/view/book/edcoll/9789004333901/B9789004333901-s011.xml. doi:https://doi.org/10.1163/9789004333901_011.

[6] A. Goyal, V. Gupta, M. Kumar, Recent named entity recognition and classification techniques: A systematic review, Computer Science Review 29 (2018) 21–43. URL: https://www.sciencedirect.com/science/article/pii/S1574013717302782. doi:https://doi.org/10.1016/j.cosrev.2018.06.001.

[7] M. Marrero, J. Urbano, S. Sánchez-Cuadrado, J. Morato, J. M. Gómez-Berbís, Named entity recognition: Fallacies, challenges and opportunities, Computer Standards Interfaces 35 (2013) 482–489. URL: https://www.sciencedirect.com/science/article/pii/S0920548912001080. doi:https://doi.org/10.1016/j.csi.2012.09.004.

[8] B. Magnini, E. Pianta, C. Girardi, M. Negri, L. Romano, M. Speranza, V. Bartalesi Lenzi, R. Sprugnoli, I-CAB: the Italian content annotation bank, in: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), European Language Resources Association (ELRA), Genoa, Italy, 2006. URL: http://www.lrec-conf.org/proceedings/lrec2006/pdf/518_pdf.pdf.

[9] P. Basile, A. Caputo, A. Gentile, G. Rizzo, Overview of the evalita 2016 named entity recognition and linking in italian tweets (neel-it) task, 2016.

[10] V. Bartalesi Lenzi, M. Speranza, R. Sprugnoli, Named entity recognition on transcribed broadcast news at evalita 2011, in: B. Magnini, F. Cutugno, M. Falcone, E. Pianta (Eds.), Evaluation of Natural Language and Speech Tools for Italian, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 86–97.

[11] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, EVALITA 2023: Overview of the 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[12] T. Paccosi, A. Palmero Aprosio, KIND: an Italian multi-domain dataset for named entity recognition, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 501–507. URL: https://aclanthology.org/2022.lrec-1.52.

[13] S. Tonelli, R. Sprugnoli, G. Moretti, Prendo la parola in questo consesso mondiale: A multi-genre 20th century corpus in the political domain, in: CLiC-it, 2019.

[14] J. D. Lafferty, A. McCallum, F. C. N. Pereira, Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, in: Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001, pp. 282–289. URL: http://dl.acm.org/citation.cfm?id=645530.655813.

[15] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[16] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the Association for Computational Linguistics 5 (2017) 135–146.

[17] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, Dbpedia: A nucleus for a web of open data, in: K. Aberer, K.-S. Choi, N. Noy, D. Allemang, K.-I. Lee, L. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, P. Cudré-Mauroux (Eds.), The Semantic Web, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007, pp. 722–735.

[18] C. D. Hromei, D. Croce, V. Basile, B. Roberto, ExtremITA@EVALITA2023: Multi-Task Sustainable Scaling to Large Language Models as its Extreme, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[19] G. Sarti, M. Nissim, IT5: Large-scale Text-to-text Pretraining for Italian Language Understanding and Generation, ArXiv preprint 2203.03759 (2022). URL: https://arxiv.org/abs/2203.03759.

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

[21] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, arXiv preprint arXiv:2106.09685 (2021).

[22] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T. B. Hashimoto, Stanford alpaca: An instruction-following llama model, https://github.com/tatsu-lab/stanford_alpaca, 2023.

CLinkaRT at EVALITA 2023: Overview of the Task on Linking a Lab Result to its Test Event in the Clinical Domain

Begoña Altuna[1,*], Goutham Karunakaran[2], Alberto Lavelli[3], Bernardo Magnini[3], Manuela Speranza[3] and Roberto Zanoli[3]

[1]HiTZ Center - Ixa, University of the Basque Country UPV/EHU, Manuel Lardizabal 1, 20018 Donostia, Spain
[2]Università di Trento, Trento, Italy
[3]Fondazione Bruno Kessler, Via Sommarive 18, 38123 Povo, Italy

Abstract

CLinkaRT at EVALITA 2023 is a relation extraction task based on clinical cases taken from the E3C corpus, i.e. Italian written documents reporting statements of a clinical practice. The task consists in identifying clinical results and measures and linking them to the laboratory tests and measurements from which they were obtained. Three teams participated in the task and various supervised machine learning models, both traditional and based on deep learning, were evaluated. In this evaluation, the deep learning models outperformed the traditional ones. Interestingly, none of the teams explored the use of few-shot language modeling. However, the fact that the supervised models significantly outperformed the task baselines implementing few-shot learning shows the crucial role still played by the availability of annotated training data.

Keywords

Relation Extraction, Clinical NLP, Named Entity Recogniton, Supervised Learning

1. Introduction and Motivation

There is a growing interest in processing clinical data for tasks of public interest, such as clinical decision making [1] or monitoring of the health status of a country [2]. While for this purpose large amounts of structured data are needed, the reality is that most clinical data are stored as free unstructured clinical texts. Hence, the ability of extracting information directly from natural language texts and to increase the volume of databases and structured datasets, such as MIMIC-III [3], is crucial.

Having these goals into account, scholars have developed a series of resources for information extraction from clinical texts. Clinical information extraction efforts have often given priority to the identification of diseases [4] or events [5]. As far as the extraction of relations from clinical texts is concerned, previous work has focused on concept normalization [6] and temporal relations [7], among others. Laboratory tests and measurements and their results have been given little attention [8], although they provide interesting information on the patients' status at a certain time of the development of a disorder and are crucial to choose the right diagnosis. From a more technical point of view, processing laboratory tests and their results also brings up a new perspective on the treatment of data, since it requires interpreting numeric values and ranges and therefore can not be handled as a common named entity recognition task [9]. In this context, the CLinkaRT task (LINKing A Result to its Test in the CLINnical domain) in EVALITA 2023 [10] provides an opportunity to evaluate different Natural Language Processing approaches and does this with a focus on Italian, a less explored language than English.

2. Task Description

The CLinkaRT task consists in identifying textual mentions of both laboratory tests and measurements in a clinical narrative, and then linking these to their respective results. Clinical narratives (or clinical cases) are documents reporting statements of a clinical practice, presenting the reason for a clinical visit, the description of physical exams, the assessment of the patient's situation and the diagnosis, as well as the following treatments. Laboratory tests and measurements are commonly done as part of this process and are typically documented in clinical narratives.

Figure 1 presents an excerpt of a clinical case where laboratory tests have been marked in bold[1] and their results in italics.

[1]Note that the head of the mention is capitalized.

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
*Corresponding author.
†These authors contributed equally.
begona.altuna@ehu.eus (B. Altuna); goutham.karunakaran@studenti.unitn.it (G. Karunakaran); lavelli@fbk.eu (A. Lavelli); magnini@fbk.eu (B. Magnini); manspera@fbk.eu (M. Speranza); zanoli@fbk.eu (R. Zanoli)
0000-0002-4027-2014 (B. Altuna)
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper43.pdf

Osvaldo, anni 52, ha una storia di diarrea e calo ponderale che si può far riferire a due anni prima. Non c'è storia di sanguinamento gastroenterico ed **una RICERCA di sangue occulto fecale** è risultata *negativa* su tre campioni. Ammette di averci dato dentro con l'alcol in passato, ma da diversi anni è assolutamente astinente. Ha un diabete, controllato con insulina. Sei anni prima è stato colecistectomizzato. **Gli ESAMI di laboratorio** sono *normali*, se si fa eccezione per una lieve anemia, così come *normali* sono **lo STUDIO radiologico del piccolo e del grosso intestino**.

Figure 1: A sample clinical case

In this example, we have the following Pertains-To relations that participants needed to identify between results and tests:

- negative / negativa −> the research for fecal blood / una RICERCA di sangue occulto fecale
- normal / normali −> the laboratory exams / gli ESAMI di laboratorio
- normal / normali −> the radiological study of the bowels / lo STUDIO radiologico del piccolo e del grosso intestino

3. Dataset

The CLinkaRT task is based on the Italian part of E3C, the multilingual European Clinical Case Corpus [11], a collection of clinical cases derived from different sources, such as published articles available from PubMed, and existing corpora. As such, the dataset encompasses a variety of clinical disciplines in different hospitals and a wide range of laboratory tests.

One of the three sections which make up the E3C corpus has been manually annotated with different types of information, such as:

- events (which include laboratory tests, among others), temporal expressions and temporal relations, annotated according to THYME [12], an adaptation of the TimeML framework [13];
- results of laboratory tests and measurements, marked through the RML tag (defined within the E3C project), and Pertains-To relations holding between an RML and the event it refers to;
- clinical entities (in particular diseases, syndromes, findings, signs, symptoms, etc.) listed in medical taxonomies, which is useful for tasks focusing on clinical entity recognition and analysis [14].

More specifically, the CLinkaRT task is based on two sets of data:

1. Training and development data: 83 clinical cases, for a total of 28,856 tokens and 658 Pertains-To relations. These documents correspond directly with the manually annotated section of the E3C corpus and have then been revised for the task;
2. Test data: 80 clinical cases taken from E3C and specifically annotated for the task, for a total of 26,437 tokens and 612 Pertains-To relations.

3.1. Annotation

Among all the annotations foreseen by the E3C project, the data used for CLinkaRT contain the following annotations:

- Laboratory test and measurement EVENTS: they include medical procedures in which parts of the body or bodily substances (blood, urine, etc.) are analyzed, as well as different acts of measuring, such as measuring patients' physical features (e.g. height and weight) or the size of a lesion or mass.
- RMLs are the results of lab tests and measurements; they can consist of a text string (e.g. normal / normali) but more often contain numerical values, typically followed by a unit of measure (e.g. 7,5 g/dl);
- Pertains-To relations connecting an RML (the source) to the relevant EVENT (the target). Pertains-To relations can be one-to-one, one-to-many and many-to-one.

In the example below we have two Pertains-To relations between two EVENTS, i.e. a laboratory test (*protidemia totale*) and a measurement (*peso*), and their results (RMLs), i.e. 4,5 g/dl and 19 Kg respectively.

Peso corporeo di 19 Kg, protidemia totale 4,5 g/dl / Body weight of 19 Kg, total protidemia 4,5 g/dl
19 Kg −> Peso
4,5 g/dl −> protidemia

Both RMLs and test and measurements EVENTS are marked as strings of text; notice, however, that tests and measurements belong to the TimeML category EVENT and are therefore marked by their syntactic head only (i.e. strictly one token only) while RMLs, as defined within the E3C project, are marked by a whole syntactic chunk (one or more tokens).

3.2. Inter Annotator Agreement

All the data used for the task have been (manually) annotated by expert computational linguists and inter-annotator agreement has been assessed on ten documents, which have been annotated by two annotators independently. On average, each annotator has identified 111 relations.

100001|t|Osvaldo, anni 52, ha una storia di diarrea e calo ponderale che si può far riferire a due anni prima. Non c'è storia di sanguinamento gastroenterico ed **una RICERCA di sangue occulto fecale** è risultata *negativa* su tre campioni. Ammette di averci dato dentro con l'alcol in passato, ma da diversi anni è assolutamente astinente. Ha un diabete, controllato con insulina. Sei anni prima è stato colecistectomizzato. **Gli ESAMI di laboratorio** sono *normali*, se si fa eccezione per una lieve anemia, così come *normali* sono **lo STUDIO radiologico del piccolo e del grosso intestino.**

100001	REL	201-209	negativa	156-163	ricerca
100001	REL	442-449	normali	416-421	esami
100001	REL	502-509	normali	518-524	studio

Figure 2: Sample of an annotated clinical case

The resulting Dice's coefficient [15] is 0.87, which is quite high given that agreement between annotators is only considered as such when there is a complete overlap in the spans of the source and the target (exact match). The high agreement between annotators ensures that annotations throughout the whole dataset are consistent. More specifically, the inter-annotator agreement is particularly high when numerical values are present in the RMLs (it reaches 0.92 in terms of Dice's coefficient), while it is slightly lower (Dice=0.84) in the case of RMLs without numerical values.

3.3. Data Distribution Format

The annotated data have been provided to the participants in a format that is in an adaptation of the PubTator format (see an example in Figure 2). It consists of a straightforward tab-delimited text file, where every document in the dataset is in a new line preceded by the DOCID and the `|t|` marker. A space line is used as an indicator of the end of the document, followed by the annotated relations: every relation is in a separate line and is represented as an ordered pair, as in (RML −> EVENT), and each string is represented by its start and end character offsets.

4. Baselines

To improve the assessment of participant systems' performance, supervised and unsupervised baselines have been used for comparative analysis. These baselines have been made available through the GitLab repository.[2]

The supervised baseline was assessed using two different approaches.

The first approach is based on vocabulary-transfer from training to testing (voc. tran.). In this approach, a system is used to recognize textual references to laboratory tests and measurements present in the test set using the entities found in the training set. Additionally, regular expressions derived from the training data are used to recognize various result entities that pertain to measurements, typically represented by values. To establish relationships between the recognized entities, a relation is created for each pair of laboratory test/measurement and result entities that co-occur together within the same sentence.

The second approach relies on a fine-tuned multilingual BERT model[3] trained on textual mentions involved in relations within the training data. The implementation of this model has been carried out using the SimpleTransformer library.[4] The model is capable of recognizing both textual references to laboratory tests and measurements and their results.

Peso corporeo di 19 Kg, protidemia totale 4,5 g/dl / Body weight of 19 Kg, total protidemia 4,5 g/dl
19 Kg −> Peso
4,5 g/dl −> protidemia

In the example above the implemented model identifies the following mentions, using the IOB annotation where test events are represented as TST and results as RML:

```
Peso [B-TST] corporeo di 19 [B-RML] Kg
[I-RML], protidemia [B-TST] totale 4,5
[B-RML] g/dl [I-RML])
```

Subsequently, an additional multilingual BERT model (configured similarly to the previous BERT model) was fine-tuned on the annotated relations within the training data to extract the relationships between the recognized laboratory tests and their results in the test data. Concerning the training data, both positive and negative examples were generated for sentences containing at least one laboratory test/measurement and one result entity. For each generated example, the entities in the relationship were

[2] https://gitlab.fbk.eu/zanoli/clinkart-baseline.git

[3] model=bert-base-multilingual-cased, epochs=5, learning_rate=4e-5, batch_size=16

[4] https://simpletransformers.ai

marked by adding "[TST]" as both the prefix and suffix to the laboratory tests and measurements, while "[RML]" was used to denote the results. The number of examples generated per sentence was determined by multiplying the number of laboratory tests by the number of result entities present in the sentence.

For the test data, the examples to be classified were generated following a similar process, with the difference that instead of using the entities from the gold standard we used the predicted entities. In the case of the sentence reported above, the following examples were generated along with their corresponding model predictions (1=positive, 0=negative):

```
1 [TST]Peso[TST] corporeo di [RML]19
Kg[RML], protidemia totale 4,5 g/dl

0 [TST]Peso[TST] corporeo di 19 Kg,
protidemia totale [RML]4,5 g/dl[RML]

0 Peso corporeo di [RML]19 Kg[RML],
[TST]protidemia[TST] totale 4,5 g/dl

1 Peso corporeo di 19 Kg,
[TST]protidemia[TST] totale [RML]4,5
g/dl[RML]
```

The unsupervised baseline uses GPT and OpenAI's API (text-davinci-003). It focuses on one-shot learning, where the model receives a single example during inference through the prompt. This makes one-shot learning more similar to unsupervised learning than supervised learning. The prompt used for performing this evaluation is: *Ho un compito che è quello di estrarre menzioni di test di laboratorio e dei loro risultati da casi clinici. Ecco un esempio di testo e output: docId:100998. Nota: nell'output viene scritto prima il risultato e poi il nome del test. Sono separati da "|". Ora dammi l'output per il seguente testo.*[5] Within the prompt, docId:100998 represents the annotated document selected from the training dataset as the only example for GPT.

5. System Descriptions

Eight teams expressed their interest to participate in the task. Eventually, four teams submitted their annotated data, resulting in a total of six runs. After the evaluation phase, one team decided to withdraw so we now present the results of four runs submitted by three different teams.

[5]My task is to extract laboratory test mentions and their results from clinical cases. Here you have an example of a text and its output: docId:100998. Notice: in the output you first write the result and then the name of the test. They are separated by "|". Now give me the output for the following text.

Participants explored various (supervised) approaches, including traditional machine learning methods, as well as using BERT [16] and its derivative models, and top Large Language Models (LLMs) such as LLaMA [17]. A brief overview of each team's approach is reported below, while the corresponding results are reported in Table 1.

Simple Ideas: Unlike conventional methods that extract entities and relations separately in a pipeline, the proposed approach uses a pipeline in which first EVENTS are identified and then the Pertains-to relations are created from those. Several BERT-based models were assessed, including Italian BERT [18] and DistilBERT [19], which were pre-trained on general topics. Additionally, BioBIT and MedBIT-R3-plus [20] were evaluated as they were specifically pre-trained for the medical field. Among these models, MedBIT-R3-plus resulted as the best model. To optimize their performance, the models were fine-tuned on an augmented version of the original dataset. This augmentation involved the addition of new sentences derived from the original ones, wherein random words were substituted with similar words in the embedding space. This approach achieved the best results in the task and it also obtained the highest ranking in the parallel TESTLINK task at IberLEF 2023 [21]. The availability of the implemented code contributes to the reproducibility of the presented results.

ExtremITA: The team employed a unified neural model to address all the EVALITA 2023 tasks. To achieve this, they experimented with two different approaches. One approach involved fine-tuning an encoder-decoder model, specifically T5 [22] pre-trained on Italian texts. The second approach is an instruction-tuned Decoder-only model based on the LLaMA [17] foundational models. This model was initially trained on Italian translations of Alpaca [23] instruction data. In both cases, the models were fine-tuned by using the complete set of datasets provided by the EVALITA 2023 tasks. Moreover, the CLinkaRT dataset was expanded with annotated documents derived from the Spanish dataset made available in the TESTLINK task. The model built upon the LLaMA model showed strong performance across multiple tasks at EVALITA 2023, including the CLinkaRT task, where it ranked second. The implemented code has been made available.

Polimi: The team used a traditional pipeline-based approach for relation extraction. The first module focused on recognizing entities related to laboratory tests and their corresponding measurements. The module was implemented using two diverse models: CRF [24] and UmBERTo [25]. For training the CRF, a range of lexical features were used, along with external sources of knowledge like UMLS [26]. Subsequently, the second module aimed at establishing relationships between exams and results by pairing them based on proximity within the same sentence. While the CRF method obtained quite sat-

isfactory results, tokenization issues prevented any results from being obtained using UmBERTo.

6. Results

We conducted the evaluation of systems' performance using the BioCreative V CDR task[6] scorer In this evaluation, a relation prediction is considered correct if the start and end character offsets of the source and target entities, as well as their order within the relation, are all accurately predicted.

The results of the systems that participated in the task, computed using the standard Precision (Pr), Recall (Re), and F_1 measures, are presented in Table 1. As a comparative reference, we report that the F_1 measure obtained by the best system in the parallel TESTLINK task is 68.38 and 72.65 for Spanish and Basque, respectively. The results of the baselines described in Section 4 are reported in Table 2.

Team	Pr	Re	F_1
Simple Ideas-BERT	65.55	60.62	62.99
ExtremITA-LLaMA	71.10	50.65	59.16
Polimi-CRF	70.34	27.12	39.15
ExtremITA-T5	46.82	26.47	33.82

Table 1
Precision, Recall and F_1 measure obtained by the participating systems.

Baseline	Type	Pr	Re	F_1
mBERT	S	61.37	64.37	62.83
GPT	U	29.55	48.73	36.79
voc. tran.	S	29.95	31.86	30.88

Table 2
Precision, Recall and F_1 measure obtained by the supervised (S) and unsupervised (U) baselines.

We additionally evaluated systems' performance (in terms of F_1 measure) based on two different dimensions. Table 3 shows the results distinguishing two categories of relations, i.e. n-ary relations (one-to-many and many-to-one) and one-to-one relations. Table 4 presents separate results for relations involving numerical RMLs and non-numerical RMLs. Finally, Table 5 reports the accuracy of participant systems in the recognition of RMLs and EVENTs, i.e. the sources and targets of the relations.

[6] https://biocreative.bioinformatics.udel.edu/tasks/biocreative-v/track-3-cdr/

Team	n-ary	one-to-one
Simple Ideas-BERT	37.50	70.77
ExtremITA-LLaMA	30.77	65.16
Polimi-CRF	ND	46.60
ExtremITA-T5	10.10	33.38

Table 3
F_1 measure of participating systems across n-ary and one-to-one relations.

Team	non-numerical	numerical
Simple Ideas-BERT	43.03	79.09
ExtremITA-LLaMA	47.35	66.36
Polimi-CRF	18.64	49.21
ExtremITA-T5	28.20	37.76

Table 4
F_1 measure of participating systems across relations involving numerical and non-numerical RMLs.

Team	EVENT	RML
Simple Ideas-BERT	74.89	75.26
ExtremITA-LLaMA	65.69	74.63
Polimi-CRF	48.85	49.80
ExtremITA-T5	48.80	58.39

Table 5
F_1 measure per entity type.

7. Discussion

Both traditional machine learning and more recent deep learning models were tested for relation extraction. It is worth noting that all participating systems were based on supervised approaches. Additionally, every system outperformed the vocabulary transfer baseline, which represents the threshold below which systems are not expected to perform.

Surprisingly, none of the teams attempted to evaluate few-shot learning with LLMs such as GPT [27] or LLaMA [17]. However, ExtremITA did evaluate LLaMA, but instead of employing few-shot learning, they opted for a fine-tuning approach, refining the model using the available training data.

The assessment of the GPT-based baseline highlights the present understanding that few-shot learning cannot be considered a viable alternative to fine-tuning in the context of the present task. Fine-tuning, although requiring annotated data, produces significantly better results.

Despite using different pre-trained models trained on diverse domain-specific data (generic domain vs medical domain), the top-performing team (Simple Ideas), along with the second-placed team (ExtremITA) and the baseline model based on multilingual BERT (mBERT), achieved remarkably similar results.

CRF (Polimi), as the exclusive traditional machine learning algorithm involved in the task, obtained a preci-

sion (70.34) in line with that of the top models (71.10). Nevertheless, its relatively lower recall (27.12), in comparison to the recall of the best-performing models (60.62), results in moderately satisfactory outcomes in terms of F_1 score (39.15).

One team (Simple Ideas) conducted an evaluation of their pipeline-based approach in two distinct tasks: the Italian CLinkaRT task (F_1 62.99) and the parallel TESTLINK task at IberLEF 2023, focusing on Basque (F_1 72.65) and Spanish (F_1 68.38). Interestingly, this approach demonstrated superior performance results across all three languages.

Based on the outcome of our analysis of systems' performance in relation to two distinct diagonal dimensions, i.e. n-ary and one-to-one relations on one hand, and numerical and non-numerical RMLs on the other (see Tables 3 and 4, we can observe that extracting n-ary relations is more challenging than extracting one-to-one relations, which is not surprising. Moreover, the task of extracting relations involving numerical RMLs seems easier than extracting relations involving non-numerical entities, which may be correlated to the lower agreement obtained on the latter in the IAA test.

An analysis of the entities involved in the relations extracted by the participants' systems shows that recognising EVENTs seems to be generally harder than recognising RMLs (Table 5). One possible explanation for this is that EVENTs are commonly identified by their syntactic head (leaving out the other elements in the phrase) which can sometimes be quite challenging.

Participants report two key reasons for the incorrect tagging produced by their models. On one hand, BERT tokenizers struggle splitting correctly medical terms (e.g. antitrombina -> anti trombina), which leads to wrongly setting the boundaries of the annotations. In addition, the difficulty of capturing the most peripheral elements in the entity mentions has also been a cause for failing to detect correctly the entity spans. This is the case of "punte di [circa 1200 pg/ml]" or "pari a 0 o [inferiori a 1.5 mg/dl]" in which only the tokens between the brackets have been annotated by the systems.

The results obtained did not allow us to determine whether the task being examined is inherently more difficult in one language compared to other languages due to language-specific traits. Within this framework, the vocabulary transfer baseline, which is expected to provide a preliminary indication of the task's difficulty, achieves better results on the Italian CLinkaRT task (F_1 30.88) compared to the parallel TESTLINK task for Basque (F_1 23.96) and Spanish (F_1 22.10). However, the participating systems, such as the Simple Idea's system, showed contrasting results.

8. Conclusions

Extracting laboratory tests and measurements and their results from clinical narratives seems to be a challenging task in clinical information extraction. The great variety of tests and the fact that most results contain numerical values differentiate this task from most entity recognition and linking tasks. Participant systems have achieved good results but there is still room for improvement, especially as far as recall is concerned. As this was the first time that we were proposing this task, we decided to keep it strictly focused on relations between tests and their results, but in the future it might be interesting to integrate this task in a more complex information extraction effort that considers a wider range of clinical entities and relations.

Acknowledgments

This work has been partially funded by the Basque Government postdoctoral grant POS 2022 2 0024.

References

[1] K. Jain, V. Prajapati, NLP/Deep Learning Techniques in Healthcare for Decision Making, Primary Health Care 11 (2021). URL: https://www.iomcworld.org/open-access/nlpdeep-learning-techniques-in-healthcare-for-decision-making-66608.html.

[2] O. Sankoh, P. Byass, Cause-specific mortality at INDEPTH Health and Demographic Surveillance System Sites in Africa and Asia: concluding synthesis, Global health action 7 (2014). doi:10.3402/gha.v7.25590.

[3] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R. G. Mark, MIMIC-III, a freely accessible critical care database, Scientific Data 3 (2016). URL: http://www.timeml.org/publications/timeMLpubs/IWCS-v4.pdf.

[4] O. Trigueros, A. Blanco, N. Lebeña, A. Casillas, A. Pérez, Explainable ICD multi-label classification of EHRs in Spanish with convolutional attention, International Journal of Medical Informatics 157 (2022) 104615. URL: https://www.sciencedirect.com/science/article/pii/S1386505621002410. doi:https://doi.org/10.1016/j.ijmedinf.2021.104615.

[5] S. Santiso, A. Pérez, A. Casillas, Adverse Drug Reaction extraction: Tolerance to entity recognition errors and sub-domain variants, Computer Methods and Programs in Biomedicine 199 (2021) 105891. URL: https://www.sciencedirect.com/science/article/pii/S0169260720317247. doi:https://doi.org/10.1016/j.cmpb.2020.105891.

[6] D. Newman-Griffis, G. Divita, B. Desmet, A. Zirikly, C. P. Rosé, E. Fosler-Lussier, Ambiguity in medical concept normalization: An analysis of types and coverage in electronic health record datasets, Journal of the American Medical Informatics Association 28 (2020) 516–532. URL: https://doi.org/10.1093/jamia/ocaa269. doi:10.1093/jamia/ocaa269.

[7] G. Alfattni, N. Peek, G. Nenadic, Extraction of temporal relations from clinical free text: A systematic review of current approaches, Journal of Biomedical Informatics 108 (2020) 103488. URL: https://www.sciencedirect.com/science/article/pii/S1532046420301167. doi:https://doi.org/10.1016/j.jbi.2020.103488.

[8] T. Hao, H. Liu, C. Weng, Valx: A System for Extracting and Structuring Numeric Lab Test Comparison Statements from Text, Methods of information in medicine 55 (2016) 266–75. doi:10.3414/ME15-01-0112.

[9] B. Percha, Modern Clinical Text Mining: A Guide and Review, Annual Review of Biomedical Data Science 4 (2021) 165–187. URL: https://doi.org/10.1146/annurev-biodatasci-030421-030931. doi:10.1146/annurev-biodatasci-030421-030931, pMID: 34465177.

[10] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[11] B. Magnini, B. Altuna, A. Lavelli, M. Speranza, R. Zanoli, The E3C Project: Collection and Annotation of a Multilingual Corpus of Clinical Cases, in: Proceedings of the Seventh Italian Conference on Computational Linguistics, Associazione Italiana di Linguistica Computazionale, Bologna, Italy, 2020. URL: http://ceur-ws.org/Vol-2769/paper_55.pdf.

[12] W. F. Styler, S. Bethard, S. Finan, M. Palmer, S. Pradhan, P. C. de Groen, B. Erickson, T. Miller, C. Lin, G. Savova, et al., Temporal Annotation in the Clinical Domain, Transactions of the Association for Computational Linguistics 2 (2014) 143–154. URL: http://aclweb.org/anthology/Q14-1012.

[13] J. Pustejovsky, J. M. Castaño, R. Ingria, R. Saurí, R. J. Gaizauskas, A. Setzer, G. Katz, D. R. Radev, TimeML: Robust Specification of Event and Temporal Expressions in Text, New directions in question answering 3 (2003) 28–34. URL: http://www.timeml.org/publications/timeMLpubs/IWCS-v4.pdf.

[14] R. Zanoli, A. Lavelli, D. Verdi do Amarante, D. Toti, Assessment of the E3C corpus for the recognition of disorders in clinical texts, Natural Language Engineering (2023) 1–19. doi:10.1017/S1351324923000335.

[15] L. R. Dice, Measures of the amount of ecologic association between species, Ecology 26 (1945) 297–302. URL: http://www.jstor.org/pss/1932409.

[16] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[17] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, CoRR abs/2302.13971 (2023). URL: https://doi.org/10.48550/arXiv.2302.13971. doi:10.48550/arXiv.2302.13971. arXiv:2302.13971.

[18] S. Schweter, Italian BERT and ELECTRA models. Version 1.0.1, 2020. URL: https://doi.org/10.5281/zenodo.4263142. doi:10.5281/zenodo.4263142.

[19] V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, in: 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing @ NeurIPS 2019, 2019. URL: http://arxiv.org/abs/1910.01108.

[20] T. M. Buonocore, C. Crema, A. Redolfi, R. Bellazzi, E. Parimbelli, Localising in-domain adaptation of transformer-based biomedical language models, ArXiv abs/2212.10422 (2022).

[21] B. Altuna, R. Agerri, L. Salas-Espejo, J. J. Saiz, R. Zanoli, M. Speranza, B. Magnini, A. Lavelli, G. Karunakaran, Overview of TESTLINK at IberLEF 2023: Linking Results to Clinical Laboratory Tests and Measurements, Procesamiento del Lenguaje Natural 71 (2023).

[22] G. Sarti, M. Nissim, IT5: Large-scale Text-to-text Pretraining for Italian Language Understanding and Generation, ArXiv preprint 2203.03759 (2022). URL: https://arxiv.org/abs/2203.03759.

[23] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T. B. Hashimoto, Stanford Alpaca: An Instruction-following LLaMA model, https://github.com/tatsu-lab/stanford_alpaca, 2023.

[24] A. McCallum, W. Li, Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons, in: Proceed-

ings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, 2003, pp. 188–191. URL: https://aclanthology.org/W03-0430.

[25] F. Tamburini, How "BERTology" Changed the State-of-the-Art also for Italian NLP, in: J. Monti, F. dell'Orletta, F. Tamburini (Eds.), Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021, volume 2769 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: http://ceur-ws.org/Vol-2769/paper_79.pdf.

[26] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology., Nucleic Acids Res. 32 (2004) 267–270. URL: http://dblp.uni-trier.de/db/journals/nar/nar32.html#Bodenreider04.

[27] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Simple Ideas at CLinkaRT: LeaNER and MeaNER Relation Extraction

Marius Micluța-Câmpeanu[1], Liviu P. Dinu[1,2]

[1]Faculty of Mathematics and Computer Science, University of Bucharest, Romania
[2]Human Language Technologies Research Center, University of Bucharest, Romania

Abstract

In this paper, we present our approach for performing relation extraction on clinical texts in the context of the CLinkaRT task at EVALITA 2023. Our system ranked first in this task with an F1-score of 62.99, outperforming most other submissions by a significant margin, with an increase of 6.5% over the second best score of 59.16, while also improving over the mBERT baseline of 62.83. We pursue a simple yet unexplored method to determine sentence level relations in text by relying on Named Entity Recognition models to identify the components of a relation. We apply this method to link laboratory results to their appropriate events in medical reports.

Keywords

EVALITA, CLinkaRT, named entity recognition, relation extraction, transformers

1. Introduction

The availability of vast quantities of textual data in the biomedical domain from digital repositories like PubMed Central has led to the development of highly specialized resources and language models [1]. Nonetheless, most of these efforts have been focused on English, while other less-resourced languages were largely neglected due to lack of available datasets.

The typical approach for downstream tasks in these languages is to resort to multilingual models, such as mBERT [2]. The rising need for pretrained models in languages other than English for biomedical applications materialized in the past few years with the advent of BioBIT/MedBIT for Italian [3] and similar models for other lower-resource languages: Spanish [4], Turkish [5] and French [6].

In the context of creating better systems for Italian, the CLinkaRT shared task [7] at EVALITA 2023 [8] challenges participants to detect laboratory measurements and tests from clinical records in order to associate them with their corresponding results. The relevance of developing and improving relation extraction tasks is highlighted in the literature, since it provides the underlying core elements for creating advanced biomedical text mining systems. Some examples include discovering interactions between drugs, adverse effects, genes, chemicals and diseases; predicting inappropriate emergency room visits; generating educational documents; building interaction networks and unveiling effective treatment methods for complex symptoms [9, 10].

The dataset for CLinkaRT [7] is adapted from the Italian part of the E3C Corpus, a collection of clinical narratives in several European languages [11]. The relationship annotations for this task are given in a similar format to PubTator, with each relation pair on a separate line. A pair is given by the entity mentions involved in the form of start and end offsets for sources (RML entities or results) and targets (EVENT entities or test events). A source may be a multi-token entity, while a target is always a single token. These tokenizations are provided together with the dataset to minimize evaluation mismatches.

In the following sections, we describe our team's solution to perform relation extraction in the CLinkaRT task.

2. System description

2.1. Overview

Our proposed method for this challenge is comprised of two consecutive Named Entity Recognition (NER) models that identify sources and targets for each sentence, followed by linking the entities through post-processing to output relations in the required format. The target entities predicted by the first NER model are prefixed with a special token when fed to the second NER model, which in turn identifies source entities. Since we used a similar system for our submission in the TESTLINK twin task [12], we provide a shortened description of the implementation and focus more on experiments and findings carried specifically in the context of CLinkaRT.

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
marius.micluta-campeanu@unibuc.ro (M. Micluța-Câmpeanu); liviu.p.dinu@gmail.com (L. P. Dinu)
0000-0002-7559-6756 (L. P. Dinu)
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper44.pdf

2.2. Implementation details

The first NER model is trained to predict all target entities in a sentence. For instance, in the phrase "La creatinina oscillava tra 1,5–2 mg/dL con proteinuria sempre < 1 g die" there are two relations: "creatinina" target with "1,5–2 mg/dL" as source and "proteinuria" with "< 1 g die" as source. We begin by locating targets first because the annotations mark only the syntactic head of a target, e.g. *esami* "tests" is an appropriate target for both *esami colturali* "culture tests" and *esami ematici* "blood tests".

After determining all targets in a sentence, we transform the training examples to incorporate target locations directly in text by adding a special marker token `[T]` before each target token, which should help the second NER model find relevant source entities. This is a viable strategy to denote one-to-one, one-to-many and many-to-one relations between sources and targets, thus effectively eliminating the need of a relation classifier model. The target indicates just the syntactic head, so we do not add an end marker because it might hinder the second NER model's ability to properly learn representing adequate targets.

All relation types (one-to-one, one-to-many, many-to-one) are handled in a uniform manner. For every target in a sentence, we generate one sample with a single target marker `[T]`. In this regard, there is no difference between a source with multiple targets and several one-to-one relations. For one target with many sources, only a single example is created. This way, we augment our training data for the second NER model.

As an example with two one-to-one relations, the sentence "La creatinina oscillava tra 1,5–2 mg/dL con proteinuria sempre < 1 g die" has two targets, "creatinina" and "proteinuria". The samples for our second NER model will be:

(1) "La `[T]` creatinina oscillava tra 1,5–2 mg/dL con proteinuria sempre < 1 g die" with only "1,5–2 mg/dL" labeled as source

(2) "La creatinina oscillava tra 1,5–2 mg/dL con `[T]` proteinuria sempre < 1 g die" with only "< 1 g die" labeled as source

In the following example, there is one source linked to three targets: "Gli esami colturali (germi comuni, BK) risultavano negativi.". The source is "negativi", while the targets are "esami", "germi" and "BK". Three sentences will be added, all with a single source to be predicted ("negativi"):

(3) "Gli `[T]` esami colturali (germi comuni, BK) risultavano negativi."

(4) "Gli esami colturali (`[T]` germi comuni, BK) risultavano negativi."

(5) "Gli esami colturali (germi comuni, `[T]` BK) risultavano negativi."

We encode annotations for both NER models with standard IOB2 tags (inside, outside, beginning) for either sources (RML entities) or targets (EVENT entities). A regular relation extraction pipeline would first employ a NER model to determine sources and targets at the same time and then apply a relation classifier on all possible source-target pairs.

With our approach, the first NER model is tasked to predict just target entities, while the second NER model is trained solely with labels for source entities. The consequence is that our models have a lower number of possible labels, determined by fewer IOB2 tags, therefore improving prediction performance. Each model has 3 tags: beggining, inside, outside. While the NER model used to predict targets only denotes the target head, we still need "inside" tags due to sub-word splitting required by transformer models. Contrast this with a traditional pipeline that would have a NER model with 5 tags (2 beginning tags, 2 inside tags, 1 outside tag) followed by a relation classifier model.

2.3. Data augmentation

The CLinkaRT training dataset consists of 83 Italian documents with 658 annotated relations. Due to the limited number of examples, we decide to augment the initial dataset with contextual word embeddings using the `nlpaug` library [13]. For this process, we replace random words with other similar words in the embedding space, except for labeled tokens, since there is a risk of injecting noisy labels.

We preserve the annotated entities and the target marker `[T]` in the augmented examples, ignoring sentences with 9 words or less. Samples that are not identical in terms of word count are discarded because the original labels would be misplaced.

Given that one of our main concern is finding laboratory tests and measurements, many of these entities are numeric values. To further our data augmentation, we introduce tiny changes of ± 2 for decimal values (age, year or quantities with a higher tolerance) and ± 0.1 for real values (tests or percentages). In most cases, this process should not significantly alter the labeling.

The training set has a much greater number of negative samples (examples without relations) than positive samples. We augment each example one or more times, with positive instances denoted by `in_multiplier` and negative instances by `out_multiplier`. We use the multipliers shown in Table 1, where the NER-tgt model predicts targets and the NER-src model predicts sources. The second model requires fewer auxiliary examples because the preprocessing step created additional samples for

Multiplier type	NER-tgt	NER-src
`in_multiplier`	4	2
`out_multiplier`	1	1

Table 1
Data augmentation multipliers: for each NER model and example type (labeled/unlabeled), we repeat the augmentation one or more times

System	Precision	Recall	F1-score
vocab transfer	29.95	31.83	30.88
mBERT	61.37	**64.37**	62.83
Second best team	**71.10**	50.65	59.16
Our team	65.55	60.62	**62.99**

Table 2
Baselines and best systems for linking measurements to results in clinical documents

sentences with more target entities.

The bottleneck of this augmentation process is the library call that executes the transformation. Considering that the operation runs on GPU, it should be natural to attempt to speed up this step by augmenting several examples in parallel. While the `nlpaug` library has an API that allows augmenting multiple sentences at once and at first it appeared to work on a few samples in the train set, a significant number of augmented examples constructed by `nlpaug` turned out to be empty sentences due to limitations or issues of this library. The batch implementation required a bit of effort due to the need to apply different multipliers. Since this attempted optimization did not succeed, we resumed augmenting examples one by one.

2.4. Model training and inference

We implement our NER models as standard token classifiers with the help of HuggingFace Transformers library [14]. We perform fine-tuning on a model pretrained on Italian medical textbooks, web-crawled data and translated English PubMed abstracts, available as `IVN-RIN/medBIT-r3-plus` on HuggingFace Hub [3].

All of our training experiments are carried out by mostly preserving default parameters: AdamW optimizer with $5e^{-5}$ learning rate with linear decay and no warmup steps, $1e^{-2}$ weight decay, 8 samples per batch trained for 4 epochs, with 10% examples held out for validation. Both models are trained independently using gold labels, with the second NER model (NER-src) receiving the target marker tokens `[T]` from these gold annotations.

In inference mode, the models are asked to output source and target offsets with respect to the original raw text. For each sentence converted into an example, we store the offset of the first token. Since HuggingFace Datasets library employs a separate tokenization, we align the transformed concatenated sentences with the initial full texts by using spaCy [15].

3. Results

We conduct our experiments by creating a test set from the training set with a 90/10 split in order to simulate the final test set, switching to 10-fold cross-validation for selecting appropriate values for some of the parameters regarding training and augmentation. Although the models are trained at the sentence level, this split is by document id so we do not overestimate the model performance on unseen examples.

The main results for relation extraction in the CLinkaRT task are displayed in Table 2. Our team obtains the first place across all teams, with an F1-score of 62.99, an improvement of 6.5% in F1-score over the second best competing team of 59.16 and a slight increase over the mBERT baseline with a score of 62.83. We also achieve the highest recall of 60.62 among other participants, with the second best score of 50.65, while the mBERT baseline has a recall of 64.37. We improve the baseline precision of 61.37 with a 6.8% increase, reaching a score of 65.55.

We report the performance of our system on the validation set averaged across 10 folds together with the official results. We used cross-validation to carry out parameter and model selection. Besides these 10% reserved examples for testing, the models also set aside 10% of the remaining examples for validation and hyperparameter tuning. In spite of these efforts, we notice a possible tendency of overfit. One explanation for this phenomenon is the small size of the model validation set, with too few samples to properly adjust the parameters when training. Another related explanation is given by the high variation between some of the folds, with half of the folds obtaining F1-scores over 82%, while the other folds consistently scored lower, between 73% and 77% F1-score. Even so, it might simply be the case that the test set is intentionally constructed with novel situations to determine the performance on unseen data more accurately, which would justify the gap between test and validation.

Outside the evaluation window, we repeated the inference process a second time on the test set keeping the same parameters and achieved 64.09 F1-score, showing that our approach can outrun the other systems by a greater margin than in the official results. Still, this variation is caused by the nondeterministic nature of transformer networks. We plan to analyze the extent of this variation and to limit the randomness of our system.

System	Pr	Re	F1-score
MedBIT-R3-plus	81.71	79.06	80.20
MedBIT-R3-plus (no aug)	64.55	68.99	66.41
Italian BERT	71.22	69.09	69.99
DistilBERT	79.56	71.16	75.04
BioBIT	81.03	75.08	77.81

Table 3
Results on the development set to determine the transformer model to be used in experiments for parameter tuning

4. Discussion

In this section, we present some observations regarding the design choices of our implementation and conduct an error analysis.

The data augmentation process has three main parameters: the minimum number of words that should be replaced in a sentence, `min_aug`, and the two multipliers for positive and negative examples defined earlier as `in_multiplier` and `out_multiplier`. We pick values between 3 and 6 for `min_aug`, based on the number of failed replacements, fixing the value at 4 words. The reasoning behind this decision is that the augmentation library is sometimes unable to adequately generate valid examples due to misplaced or missing words, so the gold labels cannot be applied, in turn leading to fewer examples in the train set.

The multipliers are selected by cross-validation, stopping early in case of unsatisfactory results on the first folds. For `out_multiplier`, we vary this parameter between 0 and 2 for both NER models, while for `in_multiplier` we use values in the range 1–5 for NER-tgt and 1–4 for NER-src. Our experiments confirm that augmentation is also needed for negative samples. This step has a significant impact in our system, boosting the score on the validation set with over 20 percentage points in F1-score. As it would be expected, adding too many examples by using larger multipliers eventually induces overfit.

The main drawback of augmentation is the slow data generation. As we mentioned earlier, `nlpaug` runs sequentially, so we had to make educated guesses of what combinations of parameters to include in our experiments.

For most of our experiments, we rely on the model called MedBIT-R3-plus [3] accessible on HuggingFace Hub. In order to determine if this is the right choice, we briefly examine the effectiveness of other transformer models. We consider three alternative options: Italian BERT [16], the multilingual version of DistilBERT [17] and the BioBIT model trained only on medical textbooks [3]. Undoubtedly, Italian BERT is less suitable for this task, with a substantial drop in F1-score of 14 percentage points. The results of the multilingual model are averaged only over the first 5 folds (out of 10 folds) because the training process takes much longer. We believe the increased training time is not justified given that a much smaller model can achieve better results. For this reason, we do not conduct additional experiments with multilingual transformers. Concerning the BioBIT model, it offers slightly worse results than the MedBIT-R3-plus counterpart, as noted likewise in the original paper [3]. The results are summarized in Table 3, using data augmentation for all variants except where noted otherwise. Due to time constraints, we did not run additional experiments with these models.

In addition to typical false positives and false negatives, we observe types of errors that show the system is on the right track, but fails to output the exact offsets in the reference file.

There are a couple of incomplete entity spans:

(6) The true source is “pari 0 inferiori a 1.5 mg/dl” and the predicted source is “inferiori a 1.5 mg/dl”. Other similar examples: true source is “fino a 12.8 mg/dL”, predicted source is “12.8 mg/dL”; true source is “punte di circa 1200 pg/mL”, predicted source is “circa 1200 pg/mL”.

(7) The true target is “antitrombina”, while the predicted target is “anti”

The first situation appears due to modifying comparatives not being present or being scarcely existent in the training data, which one could solve with additional examples or through careful augmentation. The second issue seems to be a defect on our side that can be handled in post-processing by inspecting the initial tokenization.

Another common mistake is the prediction of one relation instead of two (or vice versa) in the case of intervals, which we explain by ambiguities in the training set. For example, our system outputs “1.9 – 2.5 mg/dl” linked with “creatininemia”, but there are two expected relations: “1.9” linked with “creatininemia” and “2.5 mg/dl” linked with “creatininemia”. Conversely, our system detects two relations, “sostanzialmente” linked with “obiettività” and “nei limiti di norma” linked with “obiettività”, while there is only one true relation, “sostanzialmente nei limiti di norma” linked with “obiettività”.

Lastly, a challenging facet of this task is the presence of reference values for some tests, which are picked up by our model, although they are not found as gold labels because they do not represent test results. Future work in this direction should find means to distinguish between reference values and actual measurements and test values.

5. Conclusion and future work

In this paper, we detailed our contribution in the CLinkaRT task [7] at EVALITA 2023 [8], demonstrating that intuitive solutions yield competitive results. Our proposed approach achieves the best F1-score among other systems in the task of correlating laboratory tests and measurements with their results, with a 6.5% improvement in F1-score over the second best contestant.

We present a straightforward strategy to extract sentence-level relations based on two plain NER models, illustrating the learning capabilities of transformer networks to solve challenging tasks with the help of special tokens. We intend to further explore this direction since NER models are well established and usually require fewer resources than alternative relation extraction (RE) models. The presented method is not limited to the clinical domain and it can be easily applied in other contexts, with the added benefit of shorter development cycles. In certain domains and applications, the overhead of a generic RE model may be unjustified if the relations in question are simple enough.

Data augmentation is a valuable, but underused technique in natural language processing contexts. We look forward to enhancing the augmentation procedure to account for in-domain information. Another area we believe to be worth pursuing is the handling of numeric values and ranges, either by finding a way to inject fuzzy intervals or by masking these values altogether, therefore simplifying the initial problem.

Our system implementation is available at https://gitlab.com/marius.micluta-campeanu/testlink-clinkart-2023 to encourage an open environment for future work.

References

[1] I. Beltagy, K. Lo, A. Cohan, SciBERT: A pretrained language model for scientific text, 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3615–3620. URL: https://aclanthology.org/D19-1371. doi:10.18653/v1/D19-1371.

[2] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[3] T. M. Buonocore, C. Crema, A. Redolfi, R. Bellazzi, E. Parimbelli, Localising In-Domain Adaptation of Transformer-Based Biomedical Language Models, 2022. arXiv:2212.10422.

[4] C. P. Carrino, J. Armengol-Estapé, A. Gutiérrez-Fandiño, J. Llop-Palao, M. Pàmies, A. Gonzalez-Agirre, M. Villegas, Biomedical and Clinical Language Models for Spanish: On the Benefits of Domain-Specific Pretraining in a Mid-Resource Scenario, 2021. arXiv:2109.03570.

[5] H. Türkmen, O. Dikenelli, C. Eraslan, M. C. Çalli, S. S. Ozbek, Developing Pretrained Language Models for Turkish Biomedical Domain, in: 2022 IEEE 10th International Conference on Healthcare Informatics (ICHI), IEEE, 2022, pp. 597–598.

[6] Y. Labrak, A. Bazoge, R. Dufour, M. Rouvier, E. Morin, B. Daille, P.-A. Gourraud, DrBERT: A Robust Pre-trained Model in French for Biomedical and Clinical domains, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 16207–16221. URL: https://aclanthology.org/2023.acl-long.896.

[7] B. Altuna, G. Karunakaran, A. Lavelli, B. Magnini, M. Speranza, R. Zanoli, CLinkaRT at EVALITA 2023: Overview of the Task on Linking a Lab Result to its Test Event in the Clinical Domain, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[8] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, EVALITA 2023: Overview of the 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[9] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, H. Liu, Clinical information extraction applications: A literature review, Journal of Biomedical Informatics 77 (2018) 34–49. URL: https://www.sciencedirect.com/science/article/pii/S1532046417302563. doi:https://doi.org/10.1016/j.jbi.2017.11.011.

[10] N. Perera, M. Dehmer, F. Emmert-Streib, Named Entity Recognition and Relation Detection for Biomedical Information Extraction, Frontiers in Cell and Developmental Biology 8 (2020). URL: https://www.frontiersin.org/articles/10.3389/fcell.2020.00673.

doi:10.3389/fcell.2020.00673.
[11] B. Magnini, B. Altuna, A. Lavelli, M. Speranza, R. Zanoli, The E3C Project: Collection and Annotation of a Multilingual Corpus of Clinical Cases, in: J. Monti, F. Dell'Orletta, F. Tamburini (Eds.), Proceedings of the Seventh Italian Conference on Computational Linguistics, volume 2769 of *CLiC-It*, CEUR-WS, Milan Italy, 2020, pp. 422–431.
[12] B. Altuna, R. Agerri, L. Salas-Espejo, J. J. Saiz, A. Lavelli, B. Magnini, M. Speranza, R. Zanoli, G. Karunakaran, Overview of TESTLINK at IberLEF 2023: Linking Results to Clinical Laboratory Tests and Measurements, Procesamiento del Lenguaje Natural 71 (2023).
[13] E. Ma, NLP Augmentation, https://github.com/makcedward/nlpaug, 2019.
[14] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-Art Natural Language Processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.
[15] I. Montani, M. Honnibal, M. Honnibal, S. V. Landeghem, A. Boyd, H. Peters, P. O. McCann, jim geovedi, J. O'Regan, M. Samsonov, G. Orosz, D. de Kok, M. Blättermann, D. Altinok, S. L. Kristiansen, M. Kannan, R. Mitsch, R. Bournhonesque, Edward, L. Miranda, P. Baumgartner, R. Hudson, E. Bot, Roman, L. Fiedler, R. Daniels, W. Phatthiyaphaibun, G. Howard, Y. Tamura, spaCy: Industrial-strength Natural Language Processing in Python, 2023. URL: https://doi.org/10.5281/zenodo.7715077. doi:10.5281/zenodo.7715077.
[16] S. Schweter, Italian BERT and ELECTRA models, 2020. URL: https://doi.org/10.5281/zenodo.4263142. doi:10.5281/zenodo.4263142.
[17] V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, ArXiv abs/1910.01108 (2019).

Polimi at CLinkaRT: a Conditional Random Field vs a BERT-based approach

Vittorio Torri[1], Francesca Ieva[1,2]

[1]MOX - Modelling and Scientific Computing Lab, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milan, Italy
[2]HDS - Health Data Science Centre, Human Technopole, Viale Rita Levi-Montalcini 1, Milan, Italy

Abstract

In the context of the EVALITA 2023 challenge, we present the models we have developed for the CLinkaRT task, which aims to identify medical examinations and their corresponding results in Italian clinical documents. We propose two distinct approaches: one utilising a Conditional Random Field (CRF), a probabilistic graphical model traditionally used for Named Entity Recognition, and the other based on BERT, the transformer-based model that is currently state-of-the-art for many Natural Language Processing tasks. Both models incorporate external knowledge from publicly available medical resources and are enhanced with heuristic rules to establish associations between exams and results. Our comparative analysis elects the CRF-based model as the winner, securing the third position in the competition ranking, but the BERT-based model demonstrated competitive performance.

Keywords

Natural Language Processing, Named Entity Recognition, Clinical documents,

1. Introduction

The widespread adoption of Electronic Health Records (EHR) has led to a significant transformation in healthcare data collection, allowing for the accumulation of extensive patient information. However, a considerable portion of this data remains unstructured, posing challenges to its utilisation in statistical analyses. Within EHR systems, vast amounts of textual data, such as clinical notes, reports, and discharge summaries, are stored, containing valuable patient history that often lacks in traditional databases. In recent years, Natural Language Processing (NLP) advancements have opened up possibilities for extracting structured data from text. However, numerous challenges persist, particularly in specialised domains like medicine [1] and when dealing with languages other than English [2].

This paper presents the models we have developed for the CLinkaRT task [3] as part of the EVALITA 2023 challenge [4]. The task entails identifying pairs of medical examinations and their corresponding results within Italian clinical documents. To accomplish this, a subset of the Italian section of the E3C corpus [5], annotated by the task organisers, was provided as the training set.

Our first system is based on a Conditional Random Field (CRF), a probabilistic graphical model that has been widely used for Named Entity Recognition (NER) [6]. NER is an NLP task that involves identifying and categorizing specific types of entities within a given text. In our case, we apply NER to recognize examination names and their corresponding results. This model is enhanced by incorporating external knowledge from additional resources and employing rules to associate each examination with its result. We compare it with an approach based on BERT, the more recent transformer-based neural network that is currently state-of-the-art for many NLP tasks [7]. In this case, we fine-tune the latest Italian version of BERT, Umberto [8], using the E3C corpus. To exploit the entire corpus, we automatically translated documents that are in languages other than Italian. Subsequently, this fine-tuned BERT model undergoes training for token classification using the annotated training set provided for the challenge, incorporating a linear classification layer.

Both models demonstrated discrete performances in the NER tasks of identifying examinations and results, while the figures were lower for the actual CLinkaRT task, which involves associating examinations with their corresponding results. The CRF-based model achieved the best results, particularly due to higher recall on examination names and higher precision on examination results, achieving the third position in the final ranking.

The code of our models is available on GitHub[1].

The rest of this paper is structured as follows: Section 2 provides an analysis of related works, Section 3 discusses the dataset used for the task, Section 4 presents a detailed description of our system, Section 5 reports the obtained results, and Section 6 provides a comprehensive discussion on our findings.

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
vittorio.torri@polimi.it (V. Torri); francesca.ieva@polimi.it (F. Ieva)
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

[1]https://github.com/vittot/CLinkaRT-2023-Polimi

CEUR-WS.org/Vol-3473/paper45.pdf

2. Related works

The application of Natural Language Processing (NLP) techniques to Italian medical documents has been relatively limited. However, a few studies have addressed tasks relevant to this challenge. Viani et al. [9] focused on identifying various entities within Italian cardiology reports, including specific examination results and drug prescriptions. Their approach involved a pipeline utilising dictionary lookup and an ontology with regular expressions linked to concepts. They developed and evaluated their methodology using a dataset of 5400 reports. In a subsequent study [10], a supervised learning approach based on recurrent neural networks was employed to extract events from a smaller dataset of 75 cardiology reports, encompassing 4300 event occurrences.

Chiaramello et al. [11] explored the mapping between relevant terms in Italian clinical notes and concepts in the Italian version of the Unified Medical Language System (UMLS) [12], including the use of the MetaMap tool [13] on Italian documents.

Another example of NER on Italian clinical data, based on recurrent neural network architecture, is [14], even if the goal, in this case, was the de-identification of clinical notes and not information extraction.

While the number of works specifically focusing on Italian documents remains limited, a more extensive body of literature exists concerning English documents. These studies predominantly employ rule-based and dictionary lookup approaches [15], conditional random fields [16], recurrent neural networks [17] and, more recently, transformer-based neural networks [18].

3. Data

The training set provided by the task organisers consists of 83 documents extracted from the Italian subset of the E3C corpus. These documents have been annotated with pairs of examination mentions and corresponding results. In particular, there are 658 pairs in the dataset, among which there are 367 unique examination names and 395 unique examination values.

The challenge ranking is based on the performance of the models on a test set consisting of 80 documents. The test set was initially released to participants without annotations.

The documents in the E3C corpus are clinical narratives originating from different sources: journal papers, admission tests for specialities in medicine, patient information leaflets for medicines, and abstracts of theses in medical science.

The CLinkaRT task poses several difficulties due to the heterogeneity of the documents and the small size of the training set. Previous works in related areas often had access to datasets comprising thousands of documents or annotations, typically from a single source and within a specific medical domain (e.g., cardiology). In our case, the documents can cover any medical area, and the concept of examinations and their results have to be intended in a broad sense.

Table 1 provides examples of annotated sentences from the training set. Sentence #1 has been annotated with two examinations: *"fluenza"* (*"fluency"*) and *"memoria"* (*"memory"*), both with the value *"valori ai limiti della norma"* (*"values at normal limits"*). This example demonstrates that the task involves not only identifying laboratory examinations with precise numerical results but also encompasses various types of examinations where results can be expressed qualitatively. Sentence #2 has been annotated as containing an examination *"calo"* (*"loss"*) with a result of *"4 kg circa"* (*"about 4 kg"*), although it can be debated whether this qualifies as an examination.

Another element of uncertainty relates to the annotations boundaries, particularly for examination results. For instance, Sentence #3 has been annotated as having the result *"della positività"* (*"of the positivity"*) for the examination *"asCa"* while the proposition *"della"* (*"of the"*) could have been excluded from the result.

It is important to note that no specific annotation guidelines have been released, at least at the present time.

The complexities arising from document heterogeneity, different possible interpretations of examination results, and the absence of comprehensive annotation guidelines highlight the challenges involved in the CLinkaRT task.

4. Description of the system

We decomposed the task problem into three subproblems:

1. NER of examination names
2. NER of examination results
3. Linking between examination names and results

For the NER subproblems, we propose the two alternative approaches of CRF and BERT in Subsection 4.1 and 4.2, respectively, while for the linking, we propose an approach based on heuristic rules in Subsection 4.3

4.1. CRF model

The primary model we developed and used for the results submission is a Conditional Random Field (CRF). A Conditional Random Field is an undirected probabilistic graphical model widely used for Named Entity Recognition (NER). The model's random variables are divided between the observed variables $\mathbf{X}$ and the output variables $\mathbf{Y}$, and the graph models the conditional probability

ID	Sentence (ITA)	Sentence (ENG)	Annotations (ITA)	Annotations (ENG)
1	La valutazione neuropsicologica ha evidenziato deficit della memoria verbale a breve e a lungo termine, della memoria di prosa e delle funzioni prassiche e valori ai limiti della norma per la memoria visuo-spaziale e per la fluenza verbale	Neuropsychological evaluation showed deficits in short- and long-term verbal memory, prose memory, and praxic functions, and values at normal limits for visuospatial memory and verbal fluency	(fluenza,valori ai limiti nella norma) (memoria,valori ai limiti nella norma)	(fluency, values at normal limits) (memory, values at normal limits)
2	Il ragazzo manifestava da circa una settimana vomiti ripetuti accompagnati da coliche addominali, inappetenza e vistoso calo ponderale (4 kg circa in una settimana)	The boy had been experiencing repeated vomiting accompanied by abdominal colic, loss of appetite, and significant weight loss (about 4 kg in a week)	(calo, 4 kg circa)	(loss, about 4 kg)
3	Alla luce della positività degli asCa	In light of the positivity of the asCa	(asCa, della positività)	(asCa, of the positivity)

Table 1
Examples of annotated documents from the training set

$P(\mathbf{Y}|\mathbf{X})$. This conditional probability is modelled as

$$P(\mathbf{y}|\mathbf{x}) \propto exp(\sum_j \lambda_j \sum_{i=1}^{n} f_j(y_{i-1}, y_i, \mathbf{x}, i))$$

where $\mathbf{x}$ is the vector of tokens (observations) that form the sequence, $\mathbf{y}$ is the vector of labels (states) over the tokens, i is an index over the sequence tokens, n is the length of the sequence, j indexes the feature functions f_j and λ_j are the parameters to be learnt. Multiple feature functions f_j can be defined, both as state feature functions or as transition feature functions. While the first ones depend on the current label y_i and on the observed sequence $\mathbf{x}$, the latter also depends on the previous label y_{i-1}.

This task has two types of entities: *examination names* and *examination results*. It is possible to use two distinct CRFs for the two types of entities or a single one, which might be preferable as it can leverage the information obtained from predicting an *examination name* label to predict an *examination result* label, and vice versa. We considered an extensive set of internal features for the CRF model, as listed in Table 2. Different combinations of them have been tested, but the best results have been achieved with the complete set of features.

All these features are computed on the current token, the previous, and the next token.

Additionally, we incorporated features related to external knowledge sources. The first source is the UMLS vocabulary. We translated each token in the training set to English and queried the English UMLS vocabulary to obtain the list of concepts corresponding to the token, with their associated semantic types. We considered a set of binary features for the presence of the 50 most relevant semantic types and a more restricted set of features only for the presence of three specific semantic types (*Laboratory or Test Result*, *Laboratory Procedure*, *Amino Acid, Peptide, or Protein*) that are most likely associated with examination names, particularly for laboratory examinations.

The second external knowledge base we used is the official medical procedures nomenclature in Lombardy Region[2]. It is a list containing the names of all medical procedures provided by the Regional Health System in Lombardy. We considered only the categories primarily related to examinations: *Anatomy-Pathological Histology-Genetics*, *Immunohematology-Transfusion*, *Clinical Chemistry*, *Laboratory in general*, *Microbiology-Virology*. We extracted a binary feature indicating if a token is present in a processed version of this list, where we removed the most frequent words (frequency > 5).

The CRF was trained with the lbgfs gradient descent algorithm, 200 maximum iterations and regularisation coefficients $c_1 = 0.03$ and $c_2 = 0.02$.

4.2. BERT-based model

BERT is a transformer-based neural network that has achieved state-of-the-art performances in many NLP tasks. Although there are no domain-specific versions of BERT for the medical domain in Italian, there are general-domain versions, the most recent of which is Umberto.

[2]https://www.dati.lombardia.it/Sanit-/Transcodifica-Codici-prestazioni/7ugz-vcug

Feature	Details
Lowercased value of the current token	
Lemmatized lowered value of the current token	Lemmas are computed with Spacy Lemmatizer
Prefix of the current token	First three characters
Suffix of the current token	Last three characters
Upper token flag	True if the current token is uppercase
Title token flag	True if the current token is lowercase, beginning with an uppercase letter
Digit flag	True if the current token is composed of digits
Math symbol flag	True if the token is a mathematical symbol
Part of speech tag	Computed with Spacy POS Tagger
Exam abbreviation flag	True if it is an acronym of two or three letters present in one of the examination names mentioned in the training set

Table 2
Set of features considered for the CRF model (excluding those related to external knowledge sources)

	B-value	I-value	B-exam	I-exam
CRF				
Precision	0.8006 (0.107)	0.8025 (0.131)	0.6501 (0.143)	0.8127 (0.192)
Recall	0.6993 (0.121)	0.7435 (0.186)	0.3991 (0.103)	0.4398 (0.232)
F1-score	0.7424 (0.100)	0.7572 (0.147)	0.4867 (0.107)	0.5245 (0.198)
BERT				
Precision	0.7256 (0.110)	0.7826 (0.126)	0.6864 (0.077)	0.6823 (0.183)
Recall	0.7204 (0.094)	0.7779 (0.147)	0.3248 (0.076)	0.4720 (0.119)
F1-score	0.7176 (0.088)	0.7748 (0.120)	0.4354 (0.079)	0.5521 (0.130)

Table 3
10-fold CV results for NER (std dev in parenthesis)

We fine-tuned Umberto using the entire E3C corpus, including labelled and unlabelled documents in all E3C languages. Non-Italian documents were automatically translated into Italian using Google Translate's APIs. A linear token-level classification layer was added to this BERT version and trained on the annotated dataset provided for the challenge while keeping the other layers frozen.

Fine-tuning of the Umberto model over the E3C corpus involved 3 epochs of training with a learning rate of $2 \cdot 10^{-5}$ and weight decay of 0.01. The last layer was trained for 50 epochs with a learning rate of 10^{-3} and weight decay of 0.01.

4.3. Linking between exams and results

We employed the following heuristic rules to link pairs of examinations and results: each exam/result is paired with the nearest result/exam within the same sentence. If there are no available elements to pair with it, it is discarded.

5. Results

Table 3 reports 10-fold cross-validation results for both models. These results are related to the NER subtask only, and they are reported for the B-Exam, I-Exam, B-Value, and I-Value, even if there is no distinction between B and I tags in the annotations, to verify if longer entities show different performances. The NER results of the two models are comparable. They show higher precision, in particular for the examination names, for which the recall is very low. The CRF has higher precision than BERT on examination results and it has higher recall on examination names. These results are not surprising, given the limited amount of training data and the large number of possible examinations that can exist in this type of data. NER results on the test set are comparable to those obtained via cross-validation over the training set (we do not report them here due to space constraints).

The CLinkaRT task evaluation is based only on recognising pairs of examinations and results. Only the pairs that precisely matched the gold standard annotations were considered for ranking and evaluation. Precision, recall, and F1-score were computed based on this precise matching. The results on the final test set for both systems, computed with the official evaluation script, are shown in Table 5. Both are aligned with the NER results in terms of precision and lower in terms of recall. The CRF model results are the best, for both precision and recall.

A manual analysis of the results highlighted that in

System	Precision	Recall	F1-Score
CRF	70.34	27.12	39.15
BERT	68.69	22.22	33.58

Table 4
Results of the two systems on the test set for the (exam, result) pairs recognition

some cases the BERT model is capturing only part of the value, while the CRF model is typically capturing it entirely or not capturing it at all. Some examples of values captured by BERT vs the gold standard: *"39"* vs *"39%"*, *"10 ng/L"* vs *"inferiori a 10 ng/L"*, *"3.6"* vs *"3.6-0.9mg/dL"*.

Another observed aspect is that the BERT model seems to be based more on the position of the words in the sentence than on the words themselves. While this is positive, sometimes it leads to recognizing as examination names words that are nearer to the value but are not the actual name (e.g.: in *"bilirubina diretta 1,8 mg/dL" ("direct bilirubin 1,8 mg/dL")* it takes *"diretta" ("direct")* as name instead of *"bilirubina" ("bilirubin")*). On the contrary, the CRF model is more based on the words themselves, at least for exam names, as it is shown by the fact that among the features with the highest weight it has many specific exam names (7 out of the first 10 features).

6. Discussion

Our two systems performed similarly on the Named Entity Recognition (NER) task. They demonstrated reasonable results for identifying *examination results*, although there is room for improvement. However, the identification of *examination names* proved to be more challenging for both systems. This can be attributed to the limited size of the training data, which made it difficult for the models to generalise to a larger set of previously unseen examination names. Despite incorporating external knowledge resources directly into the CRF model and indirectly into the BERT-based model, they were insufficient to enhance the performance in recognising a broader range of examinations. Further investigation is necessary to explore how other data sources can be utilised for this purpose. While we were unable to find an Italian dataset specifically annotated for examination names, it may be worthwhile to investigate the use of existing annotations for *clinical entities* in the E3C corpus, selecting a subset that closely aligns with the concept of examinations. Another element which is worth to further investigating is the tokenizer: for the BERT-based model, we utilised the Umberto tokenizer, but its limitations in dealing with medical terminology might have negatively affected the performances.

Regarding linking examinations to results, our naive approach based on heuristics did not appear to be a limiting factor, considering the F1-score achieved on the (exam, result) pairs compared to the F1-score for the NER of exam names. However, it is possible to explore data-driven models for this subtask, even though the scarcity of available data presents challenges.

We strongly believe in the application of Natural Language Processing techniques to the medical domain and recognise the huge need for developing models that can effectively process Italian healthcare data. Simultaneously, it is crucial to improve the quantity and quality of annotated datasets to drive the development of such models. Challenges like this are a valuable tool for motivating the academic community to contribute to this field.

References

[1] O. G. Iroju, J. O. Olaleke, A systematic review of natural language processing in healthcare, International Journal of Information Technology and Computer Science 8 (2015) 44–50.

[2] A. Névéol, H. Dalianis, S. Velupillai, G. Savova, P. Zweigenbaum, Clinical natural language processing in languages other than English: opportunities and challenges, Journal of biomedical semantics 9 (2018) 1–13.

[3] B. Altuna, G. Karunakaran, A. Lavelli, B. Magnini, M. Speranza, R. Zanoli, CLinkaRT at EVALITA 2023: Overview of the Task on Linking a Lab Result to its Test Event in the Clinical Domain , in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[4] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, EVALITA 2023: Overview of the 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[5] B. Magnini, B. Altuna, A. Lavelli, M. Speranza, R. Zanoli, The E3C Project: Collection and Annotation of a Multilingual Corpus of Clinical Cases, in: Proceedings of the Seventh Italian Conference on Computational Linguistics, 2020.

[6] H. M. Wallach, Conditional random fields: An introduction, Technical Report, Department of CIS, University of Pennsylvania (2004) 22.

[7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: 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), 2019, pp. 4171–4186.

[8] F. Tamburini, How "BERTology" changed the state-of-the-art also for Italian NLP, Computational Linguistics CLiC-it 2020 (2020) 415.

[9] N. Viani, C. Larizza, V. Tibollo, C. Napolitano, S. G. Priori, R. Bellazzi, L. Sacchi, Information extraction from Italian medical reports: An ontology-driven approach, International journal of medical informatics 111 (2018) 140–148.

[10] N. Viani, T. A. Miller, C. Napolitano, S. G. Priori, G. K. Savova, R. Bellazzi, L. Sacchi, Supervised methods to extract clinical events from cardiology reports in italian, Journal of biomedical informatics 95 (2019) 103219.

[11] E. Chiaramello, F. Pinciroli, A. Bonalumi, A. Caroli, G. Tognola, Use of "off-the-shelf" information extraction algorithms in clinical informatics: A feasibility study of MetaMap annotation of Italian medical notes, Journal of biomedical informatics 63 (2016) 22–32.

[12] O. Bodenreider, The unified medical language system (UMLS): integrating biomedical terminology, Nucleic acids research 32 (2004) D267–D270.

[13] A. R. Aronson, Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program, in: Proceedings of the AMIA Symposium, American Medical Informatics Association, 2001, p. 17.

[14] R. Catelli, F. Gargiulo, V. Casola, G. De Pietro, H. Fujita, M. Esposito, A novel Covid-19 data set and an effective deep learning approach for the de-identification of Italian medical records, Ieee Access 9 (2021) 19097–19110.

[15] M. A. Tanenblatt, A. Coden, I. L. Sominsky, The ConceptMapper Approach to Named Entity Recognition, in: LREC, 2010, pp. 546–51.

[16] H. U. Rahman, N. Chowk, T. Hahn, R. Segall, Disease named entity recognition using conditional random fields, in: Proceedings of the 7th International Symposium on Semantic Mining in Biomedicine, 2016.

[17] A. Magge, M. Scotch, G. Gonzalez-Hernandez, Clinical NER and relation extraction using bi-char-LSTMs and random forest classifiers, in: International workshop on medication and adverse drug event detection, PMLR, 2018, pp. 25–30.

[18] M. Abadeer, Assessment of DistilBERT performance on named entity recognition task for the detection of protected health information and medical concepts, in: Proceedings of the 3rd clinical natural language processing workshop, 2020, pp. 158–167.

WiC-ITA at EVALITA2023: Overview of the EVALITA2023 Word-in-Context for ITAlian Task

Pierluigi Cassotti[1,†], Lucia Siciliani[1,*,†], Lucia C. Passaro[2,†], Maristella Gatto[3,†] and Pierpaolo Basile[1,†]

[1]Department of Computer Science, University of Bari, Italy
[2]Department of Computer Science, University of Pisa, Italy
[3]Dipartimento di Ricerca e Innovazione Umanistica, University of Bari, Italy

Abstract

WiC-ita is a shared task proposed at the EVALITA 2023 campaign. The task focuses on the meaning of words in specific contexts and has been modelled as both a binary classification and a ranking problem. Overall, 4 groups took part in both subtasks, with 9 different runs. In this report, we describe how the task was set up, we report the system results, and we discuss them.

Keywords

Word in Context, Lexical Semantics, Evaluation, Dataset

1. Introduction and motivation

Word Sense Disambiguation [1] is a Natural Language Processing task with a long history and is extremely interesting for the Computational Linguistics community. In Word Sense Disambiguation (WSD), the goal is to disambiguate each word occurrence assigning to it the correct sense from a predefined sense inventory, such as WordNet [2]. The introduction of contextualized models, such as BERT, allowing the representation of a word in different contexts, steers the research focus to new tasks, such as the Word in Context (WiC) task [3].

WSD and the WiC task are highly related: while the former models in an explicit way the relationship between the target word and its sense (taken from a predefined sense inventory), the latter reduces it to a binary task. The WiC task requires determining if a word occurring in two different sentences has the same meaning or not. In recent years, there has been a growing interest in the WiC task, demonstrated by the creation of several different resources and shared tasks covering more than 20 languages.

In general, the WiC task is of broad-scope interest, as it is not limited to specific domains and can be useful for several NLP tasks. Furthermore, the training and the evaluation on a monolingual (Italian) or cross-lingual (English-Italian) dataset is advantageous not only for the models for the Italian language. In fact, the transfer learning ability of WiC models across different languages is proven in previous works [4], where models improve their performance by training in other languages. Several initiatives have been proposed throughout the years: the first one [3] being the proposal of the WiC task, which also came along with a dataset but was limited to English. For this reason, it was followed by the XL-WiC [5] dataset which tried to tackle this issue by taking into account a total of 15 languages. Next, the MCL-WiC [4] was the first WiC dataset to introduce the Cross-lingual task. The main motivation behind this particular choice was to cover scenarios where systems have to deal with different languages simultaneously, further highlighting the importance of this task in real-world applications. With AM^{2}iCo [6], the main aim was to focus on low-resource languages and to ensure participating models must consider both the target word and the context to achieve good performances. Finally, in CoSimLex [7], the task is extended to *pairs* of words that appear in a shared context, and the goal is to determine to which degree they refer to the same concept. This is done to capture the word polysemy as well as the context-dependency of words.

Shared tasks regarding the WiC usually preserve its binary design, where the two possible outcomes for each entry are: true if the meaning of the target word changes between the two sentences/contexts and false if it does not. However, there can be some cases where it is not so simple to determine the lack or presence of semantic similarity in a discrete way. For this reason, we exploit the 4-point relatedness scale introduced by [8, 9] in the annotation process. The scale consists of 4 values, namely 4: Identical; 3: Closely Related; 2: Distantly Related; 1: Unrelated. A fifth value can be assigned (0: Cannot decide) for uncertain cases.

Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023).
*Corresponding author.
†These authors contributed equally.
pierluigi.cassotti@uniba.it (P. Cassotti); lucia.sicialini@uniba.it (L. Siciliani); lucia.passaro@unipi.it (L. C. Passaro); maristella.gatto@uniba.it (M. Gatto); pierpaolo.basile@uniba.it (P. Basile)
© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper46.pdf

Unfortunately, as often happens in the Natural Language Processing research area, some languages are more represented than others, and the WiC task makes no exception in this sense. With the WiC-ITA task at EVALITA 2023 [10], we aim to fill this gap in the literature, making openly available a resource that can undoubtedly foster novel research.

2. Task Description

The general goal of the WiC-ITA task is to establish if a word w occurring in two different sentences s_1 and s_2 has the same meaning or not. The task is modelled with two different subtasks, namely a binary classification one (Subtask 1) and a ranking one (Subtask 2). Participants were allowed to participate in one or both of the subtasks. Details and examples of annotation are available on the task website. [1]

2.1. Subtask 1: Binary Classification

Subtask 1 is structured as follows: given a word w occurring in two different sentences s_1 and s_2, the goal is to provide the sentences pair with a score determining whether w maintains the same meaning or not. Possible outcomes for this subtask are:

- 0: the word w has *not* the same meaning in the two sentences s_1 and s_2;
- 1: the word w has the same meaning in the two sentences s_1 and s_2.

2.2. Subtask 2: Ranking

Subtask 2 is structured as follows: Given a word w occurring in two different sentences s_1 and s_2, the goal is to provide the sentences pair with a score indicating to which extent, in a 1-4 scale, w has the same meaning in the two sentences. The scoring system for this subtask is a continuous value where $score \in [1, 4]$. A higher score corresponds to a higher degree of semantic similarity.

3. Dataset

The creation of datasets for the WiC task usually relies on using sense inventories, such as WordNet or BabelNet [11]. More specifically, sense inventories are often exploited for selecting target words, which should exhibit polysemia, and for the generation of sentence pairs using the sense examples provided, i.e. sentences in which the target word occurs with the respective sense. After the selection of target words and the generation of sentence pairs, only a small part of these are manually annotated/-validated by human experts.

	Monolingual	**Cross-lingual**
noun	177	128
verb	65	45
adjective	49	67
adverb	11	20
	302	260

Table 1
Number of target words and number of words pairs per PoS

Differently from previous datasets, for the WiC-ITA task, we relied on sense inventories only for the target words selection stage, while we extract the list of sentence pairs from large unlabelled corpora. Moreover, human annotation is carried out for *all* the sentence pairs, thus making WiC-ITA the largest manually annotated resource for the WiC task.

In addition to this, the WiC-ITA dataset includes both monolingual (Italian) and cross-lingual (English-Italian) data.

The dataset is split into training, development, and test portions. In particular:

- the training and development set consists of annotated pairs of monolingual (Italian) sentences;
- the test set consists of annotated pairs of monolingual (Italian) sentences and annotated pairs of cross-lingual (English-Italian) sentences.

We create the monolingual datasets by selecting target words based on the number of synsets in WordNet and senses reported in Wiktionary. To achieve this, we generate a list of candidate target words for each part of speech (PoS) using lemmas from both WordNet and Wiktionary. For each lemma w, we calculate the count of WordNet synsets (wns_w) and senses reported Wiktionary (wks_w). We then compute $\min(wns_w, wks_w)$ and order all the target words in descending order.

To construct the cross-lingual dataset, we use the MultiSemCor [12], which is based on SemCor [13], the most extensive and widely used dataset for Word Sense Disambiguation. Specifically, we extracted word pairs (Italian-English) that are frequently translated in SemCor. For these word pairs, we compute the frequency of specific synsets. Then, we took the union of synsets for each word pair and computed the probability distribution over the synsets for both the Italian and English words. The Jensen–Shannon Divergence (*JSD*) is computed for each pair, and the pairs are sorted accordingly in decreasing order.

We sample the top-k words for the monolingual setting and the top-k pair of words for the cross-lingual setting according to the $\min(wns_w, wks_w)$ and the *JSD* respec-

[1] http://wic-ita.github.io/.

First annotator	Second annotator	N. examples	Spearman corr.
Annotator 2	Annotator 1	454	0.63
Annotator 3	Annotator 2	442	0.63
Annotator 5	Annotator 3	445	0.65
Annotator 6	Annotator 4	442	0.55
Annotator 6	Annotator 5	447	0.65
Annotator 7	Annotator 4	440	0.55
Annotator 8	Annotator 1	447	0.67
Annotator 10	Annotator 7	448	0.73
Annotator 10	Annotator 9	444	0.66
Annotator 11	Annotator 8	419	0.61
Annotator 11	Annotator 9	397	0.57
		4825	0.63

Table 2
Monolingual annotation statistics

First annotator	Second annotator	N. examples	Spearman corr.
Annotator 3	Annotator 2	54	0.48
Annotator 4	Annotator 1	46	0.39
Annotator 4	Annotator 3	38	0.38
Annotator 6	Annotator 2	79	0.66
Annotator 7	Annotator 2	52	0.62
Annotator 9	Annotator 5	78	0.60
Annotator 10	Annotator 1	35	0.44
Annotator 10	Annotator 5	51	0.76
Annotator 10	Annotator 8	79	0.54
Annotator 11	Annotator 7	136	0.67
Annotator 11	Annotator 9	45	0.75
		693	0.57

Table 3
Cross-lingual annotation statistics

		Class 0	Class 1
Training		806	1,999
Development	IV	236	167
	OOV	14	83
Test	IV	124	127
	OOV	126	123

Table 4
Subtask 1: number of examples for each class. IV: In-Vocabulary, OOV: Out-Of-Vocabulary

tively. The number of sampled words per PoS tag are reported in Table 1.

The monolingual and the cross-lingual sentence pairs are extracted from the itWaC and ukWaC corpora, both part of the WaCKy project [14, 15]. ukWaC is a corpus obtained by crawling the web pages under the .uk domain. It consists of more than 2 billion words, annotated with PoS tags and lemmatized using the TreeTagger tool [16]. itWaC, differently from ukWaC, is lemmatized using Morph-it! and is obtained by crawling web pages under the .it domain.

Each sentence pair extracted from the aforementioned resources has been attributed with the average score assigned by two annotators according to the 4-point relatedness scale, i.e. from 4 (Identical meaning) to 1 (Unrelated), the offsets of the target word on the respective sentences, and the lemma of the target word. Note that we only considered the Italian lemma for the cross-lingual examples, albeit providing the offsets for both languages. The annotation process is carried out using Doccano [17].

Each data point (i.e., sentence pair) is annotated by two independent annotators. Tables 2 and 3[2] show the statistics in terms of number of annotated examples and agreement (computed as the Spearman correlation) for each pair of annotators. In the monolingual setting, the Spearman correlation for the annotations consistently exceeds 0.6, with the exception of two cases. On the other hand, in the cross-lingual setting, the average correlation is lower compared to the correlation obtained in the monolingual setting. However, the correlation between annotators in the cross-lingual scenario is also computed on smaller samples, which can impact the reliability of the computed correlation.

The data points for which at least one of the annotators voted 0 (Cannot decide) were discarded from the official dataset for the sake of simplicity. The score for Subtask 2 is obtained by averaging the scores assigned by the two annotators. The ground truth labels for Subtask 1 *(binary)* were drived from the labels of Subtask 2. Specifically, we considered the data points for which the two annotators agreed, namely the case in which both annotators provided a score in the set $\{1, 2\}$ and the case in which both the annotators provided a score in the set $\{3, 4\}$. In the former case, the example was labelled with 0, while in the latter, it was labelled with 1.

The dataset is available for download on the website of the task.[3] The dataset has been constructed using available corpora. We refer to [14, 15] for the details about copyright and usage. Below, we further describe the details of the two sub-tasks.

3.1. Subtask 1: Ranking

We provide two datasets for model development:

- The training dataset which consists of 2,805 training examples. This dataset should be employed to train the model
- The development dataset which consists of 500 training examples. This dataset should be employed to evaluate the model in the training phase, e.g., tune hyper-parameters
- The monolingual test dataset which consists of 500 examples
- The cross-lingual test dataset which consists of 500 examples

The training dataset is highly unbalanced, consisting of 71.27% of positive and 28.73% of negative examples. At the same time, we provide balanced development and test datasets consisting of 50% positive and 50% of negative examples. For each In-Vocabulary target word of the

[2]The annotator groups for the two tasks are independent.
[3]https://wic-ita.github.io/data/.

development and test datasets, at least one positive and one negative example are provided in the training set. Overall statistics are reported in Table 4.

3.2. Subtask 2: Ranking

We provide four datasets for model development:

- The training dataset which consists of 2,805 training examples for which the two annotators agree (This dataset contains the same examples provided for training in Subtask 1)
- A training dataset which consists of 1,015 training examples for which the two annotators disagree
- An overall training dataset which consists of 3,820 training examples. This dataset include both instances where annotators have reached a consensus and those in disagreement
- The development dataset which consists of 500 examples (This dataset contains the same examples provided for development in Subtask 1)
- The monolingual test dataset which consists of 500 examples (This dataset contains the same examples provided for test in Subtask 1)
- The cross-lingual test dataset which consists of 500 examples (This dataset contains the same examples provided for test in Subtask 1)

4. Evaluation

The ranking of the participating systems is provided according to each subtask and test set. In other words, for each subtask, we provide the evaluation in both the monolingual and the cross-lingual setting.

4.1. Subtask 1 (Binary Classification)

Systems' predictions are evaluated against the ground truth using the macro F1-Score, i.e. we compute the F1-score for each class and we take the average of these scores to obtain the macro F1-score.

4.2. Subtask 2 (Ranking)

Systems' predictions are evaluated against the ground truth using Spearman's rank correlation. It measures the rank correlation of two variables X and Y:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (1)$$

where $d_i = R(X_i) - R(Y_i)$ is the difference between the ranks of each observation and n is the number of observations.

4.3. Baseline models

The baseline model for the task has been constructed according to [5]. It exploits the BERT architecture [18] for encoding the target sub-words. To deal with cases in which the target word is split into multiple sub-tokens, the first sub-token is considered. Differently from [5], we use as pre-trained model XLM-RoBERTa [19] and train the baseline to minimise the difference between the model prediction and the gold score computing the mean squared error.

We set the learning rate to $1e^{-5}$ and weight decay to 0. The best checkpoint over the ten epochs is selected using the development data.

The binary baseline for Subtask 1 applies the threshold $\delta = 2$ to the model predictions to obtain discrete labels.

To ensure fair reproducibility and comparisons, the evaluation scripts are available for download. [4]

5. Participants

Overall, different teams participated in the task with 9 distinct runs. We highlight below the main strategies adopted by the teams to deal with the WiC-ITA tasks.

The **BERT 4EVER** team [5] proposed three variants of a system based on BERT. The strategy behind the first model involves using the Labse pre-trained model to perform matching judgment tasks. It applies four different strategies for encoding and matching the spliced sentences, including the addition of [CLS] vectors and siamese vectors. The output probabilities of the four models are fused, with task 2 treated as a six-classification task. The second model for task 1 uses the bert-base-italian-cased pre-trained model and follows the same encoding and matching strategies as the first model. Again, the output probabilities of the four models are fused. For task 2, the Labse pre-trained model is used, and the strategies are identical to those in Model 1, but the predicted classification results are averaged. Finally, a third variant combines both the bert-base-italian-cased and Labse pre-trained models. It applies the same encoding and matching strategies as the previous models, but this time, the output probabilities of all eight models (four from each pre-trained model) are fused.

The **ExtremITA** team proposed two models fine-tuned on the EVALITA 2023 training data. The first system is based on the Large Language Model from Meta AI (LLaMA), i.e., the Italian version called Camoscio [20]. The model is pre-trained to generate text based on user instructions and fine-tuned on task-specific triples of <task, input, output> derived from the training data of EVALITA

[4] https://github.com/wic-ita/data/blob/main/evaluation.py.
[5] The team did not submit the final report.

2023 challenges. The LoRA technique for training was applied, and the model is further fine-tuned on the EVALITA 2023 training data. The second system is based on the Italian version of T5 (IT5) [21]. It underwent fine-tuning on task-specific input-output pairs derived from the training data of EVALITA 2023 challenges. The phrasal forms from the training data were used to train the model. The details of the models developed by the team are reported in [22].

The **LG** team proposed a single system based on the automatic translation of target words in different languages. Opus-MT models have been used for the translation of data into 21 languages. The words are lemmatized and aligned, and the feature vectors are created from the equivalence of the target lemma in translation. Then SVMs are used for solving tasks. PoS-Tagging and lemmatization of Italian sentences have been performed through TreeTagger[6]. Lemmatization in 21 languages has been roughly performed through Simplemma[7]. The details of the models developed by the team are reported in [23].

The models developed by **The Time-Embedding Travelers** team (afterwards mentioned as TTET) are all based on the XLM-RoBERTa-base architecture. Each model is a straightforward threshold-based classifier that utilises the conditional number of the cosine similarity or distance matrix to make predictions. The embeddings of the target word are extracted from both sentences, and pairwise similarities or distances are calculated. The threshold for classification is tuned by selecting the value that maximizes accuracy on a combined train and dev set. The final threshold for prediction is determined as the average of the threshold values obtained from multiple iterations. Model 1 and Model 2 use the last 4 layers of embeddings, while Model 3 uses embeddings from all 12 layers. The details of the models developed by the team are reported in [24].

6. Results

Table 5 reports the results referred to each subtask. Concerning the first subtask, namely the one in which participants were asked to provide a binary classification, the best results were obtained by the LG and the TTET Teams. Specifically, the LG team was ranked first on the Italian test set, while the TTET was ranked first for the Italian-English test set. The results are reported in Table 5.

With respect to the second subtask, where participants were asked to provide a ranking, the best results were obtained by the baseline for the Italian test set and by the TTET Team for the Italian-English test set. However, none of the proposed systems provided satisfactory results for the Italian test set, but the TTET team was ranked first for the Italian-English test set. The results are reported in Table 5.

[6]https://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/
[7]https://pypi.org/project/simplemma/

Team	Run	it-it	it-en
LG	LG	**0.734**	-
TTET	conditional	0.668	**0.731**
TTET	cond2	0.658	0.715
TTET	cond3	0.625	0.738
extremITA	it5	0.611	0.616
baseline		*0.594*	*0.555*
BERT 4EVER	run 3	0.556	0.492
BERT 4EVER	run 2	0.561	0.521
BERT 4EVER	run 1	0.535	0.493
extremITA	camoscio lora	0.513	0.544

Table 5
Subtask 1 (Binary Classification) Results

Team	Run	it-it	it-en
baseline		***0.569***	*0.406*
TTET	conditional	0.553	0.538
TTET	cond2	0.521	**0.548**
TTET	cond3	0.493	0.533
LG	LG	0.492	-
BERT 4EVER	run 1	0.337	0.159
BERT 4EVER	run 2	0.303	0.15
BERT 4EVER	run 3	-	-
extremITA	camoscio lora	-	-
extremITA	it5	-	-

Table 6
Subtask 2 (Ranking) Results

Table 7 presents detailed results for each system, including the classification of in-vocabulary (IV) and out-of-vocabulary (OOV) words. The aim is to evaluate the system's capability to classify words that were not part of the training data. In this regard, the LG system exhibits the highest performance in Subtask 1 for both IV and OOV words. However, in Subtask 2, only the TTET system surpasses the baseline for OOV words.

Interestingly, the performance on OOV targets shows an overall improvement. We propose that the models may have become overly specialized to the specific distribution of IV word classes during training, resulting in overfitting.

7. Conclusions

The WiC-ITA task was approached by four different teams. The results from the evaluation of four different teams' systems revealed interesting trends. While three of the systems were based on the Transformer architecture, one team developed an SVM classifier based on the output of a Machine Translation system (using the Transformer model). In the binary classification task, the

		Binary			Ranking		
Team	**Run**	**All**	**IV**	**OOV**	**All**	**IV**	**OOV**
BERT 4EVER	run 1	0.535	0.508	0.560	0.337	0.270	0.422
BERT 4EVER	run 2	0.561	0.565	0.557	0.303	0.261	0.365
BERT 4EVER	run 3	0.556	0.531	0.582	-	-	-
LG	LG	**0.734**	**0.693**	**0.775**	0.492	0.425	0.555
TTET	cond2	0.658	0.644	0.672	0.521	0.497	0.523
TTET	cond3	0.625	0.597	0.651	0.493	0.461	0.500
TTET	conditional	0.668	0.651	0.685	0.553	0.485	**0.582**
extremITA	camoscio lora	0.513	0.446	0.575	-	-	-
extremITA	it5	0.611	0.586	0.635	-	-	-
baseline		0.594	0.620	0.566	**0.569**	**0.536**	0.567

Table 7
Detailed results for subtasks and participants.

best-performing systems demonstrated a significant improvement over the baseline by 14 percentage points on the Italian test set and 17 percentage points on the English test set. However, in the ranking task, the baseline system outperformed all the proposed systems for the Italian test set, whereas the proposed systems achieved a notable enhancement of 14 percentage points over the baseline for the Italian-English test set.

For the Italian test set, the best result was achieved by the system based on SVM and Machine Translation. This team submitted results only for the monolingual task. In the English test set, the best result is obtained by the system based on the XLM-RoBERTa-base architecture. It is interesting to underline that the worst performances were obtained by the system that adopts instructions-based fine-tuning of a specific LLM for Italian. On the one hand, these results highlight the effectiveness and potential of the different systems in addressing the classification and ranking tasks for the meaning of words in context. On the other hand, the results of the competition highlight that there is still room for improvement and that the task is still far from the results obtained by similar campaigns in English.

Acknowledgments

We acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), Spoke 6 - Symbiotic AI (CUP H97G22000210007) under the NRRP MUR program funded by the NextGenerationEU.

References

[1] M. Bevilacqua, T. Pasini, A. Raganato, R. Navigli, Recent Trends in Word Sense Disambiguation: A Survey, in: Z. Zhou (Ed.), Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021, ijcai.org, 2021, pp. 4330–4338. URL: https://doi.org/10.24963/ijcai.2021/593. doi:10.24963/ijcai.2021/593.

[2] G. A. Miller, WORDNET: a lexical database for english, in: Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, USA, February 23-26, 1992, Morgan Kaufmann, 1992. URL: https://aclanthology.org/H92-1116/.

[3] M. T. Pilehvar, J. Camacho-Collados, WiC: the Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 1267–1273. URL: https://doi.org/10.18653/v1/n19-1128. doi:10.18653/v1/n19-1128.

[4] F. Martelli, N. Kalach, G. Tola, R. Navigli, SemEval-2021 Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation (MCL-WiC), in: A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot, X. Zhu (Eds.), Proceedings of the 15th International Workshop on Semantic Evaluation, SemEval@ACL/IJCNLP 2021, Virtual Event / Bangkok, Thailand, August 5-6, 2021, Association for Computational Linguistics, 2021, pp. 24–36. URL: https://doi.org/10.18653/v1/2021.semeval-1.3. doi:10.18653/v1/2021.semeval-1.3.

[5] A. Raganato, T. Pasini, J. Camacho-Collados, M. T. Pilehvar, XL-WiC: A Multilingual Benchmark for Evaluating Semantic Contextualization, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, Association for Computational Linguistics, 2020, pp. 7193–7206. URL:

https://doi.org/10.18653/v1/2020.emnlp-main.584. doi:10.18653/v1/2020.emnlp-main.584.

[6] Q. Liu, E. M. Ponti, D. McCarthy, I. Vulic, A. Korhonen, AM2iCo: Evaluating Word Meaning in Context across Low-Resource Languages with Adversarial Examples, in: M. Moens, X. Huang, L. Specia, S. W. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, Association for Computational Linguistics, 2021, pp. 7151–7162. URL: https://doi.org/10.18653/v1/2021.emnlp-main.571. doi:10.18653/v1/2021.emnlp-main.571.

[7] C. S. Armendariz, M. Purver, M. Ulcar, S. Pollak, N. Ljubesic, M. Granroth-Wilding, CoSimLex: A Resource for Evaluating Graded Word Similarity in Context, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020, European Language Resources Association, 2020, pp. 5878–5886. URL: https://aclanthology.org/2020.lrec-1.720/.

[8] D. Schlechtweg, S. S. im Walde, S. Eckmann, Diachronic Usage Relatedness (DURel): A Framework for the Annotation of Lexical Semantic Change, in: M. A. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers), Association for Computational Linguistics, 2018, pp. 169–174. URL: https://doi.org/10.18653/v1/n18-2027. doi:10.18653/v1/n18-2027.

[9] S. W. Brown, Choosing sense distinctions for WSD: psycholinguistic evidence, in: ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA, Short Papers, The Association for Computer Linguistics, 2008, pp. 249–252. URL: https://aclanthology.org/P08-2063/.

[10] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[11] R. Navigli, S. P. Ponzetto, BabelNet: Building a Very Large Multilingual Semantic Network, in: J. Hajic, S. Carberry, S. Clark (Eds.), ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden, The Association for Computer Linguistics, 2010, pp. 216–225. URL: https://aclanthology.org/P10-1023/.

[12] L. Bentivogli, E. Pianta, Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor Corpus, Natural Language Engineering 11 (2005) 247–261. doi:10.1017/S1351324905003839.

[13] G. A. Miller, C. Leacock, R. Tengi, R. Bunker, A Semantic Concordance, in: Human Language Technology: Proc. of a Workshop Held at Plainsboro, New Jersey, USA, March 21-24, 1993, Morgan Kaufmann, 1993. URL: https://aclanthology.org/H93-1061/.

[14] M. Baroni, S. Bernardini, A. Ferraresi, E. Zanchetta, The wacky wide web: a collection of very large linguistically processed web-crawled corpora, Language resources and evaluation 43 (2009) 209–226.

[15] M. Baroni, A. Kilgarriff, Large linguistically-processed web corpora for multiple languages, in: EACL'06: Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations; 2006 Apr 5-6; Trento, Italy. Stroudsburg (PA): Association for Computational Linguistics; 2006. p. 87-90, ACL (Association for Computational Linguistics), 2006.

[16] H. Schmid, Probabilistic part-of speech tagging using decision trees, in: New methods in language processing, 2013, p. 154.

[17] H. Nakayama, T. Kubo, J. Kamura, Y. Taniguchi, X. Liang, doccano: Text annotation tool for human, 2018. URL: https://github.com/doccano/doccano, software available from https://github.com/doccano/doccano.

[18] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[19] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised Cross-lingual Representation Learning at Scale, in: D. Jurafsky, J. Chai, N. Schluter, J. R. Tetreault (Eds.),

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, Association for Computational Linguistics, 2020, pp. 8440–8451. URL: https://doi.org/10.18653/v1/2020.acl-main.747. doi:10.18653/v1/2020.acl-main.747.

[20] A. Santilli, Camoscio: An italian instruction-tuned llama, https://github.com/teelinsan/camoscio, 2023.

[21] G. Sarti, M. Nissim, It5: Large-scale text-to-text pretraining for italian language understanding and generation, ArXiv preprint 2203.03759 (2022). URL: https://arxiv.org/abs/2203.03759.

[22] C. D. Hromei, D. Croce, V. Basile, R. Basili, Extremita@evalita2023: Multi-task sustainable scaling to large language models at its extreme, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[23] L. Gregori, Lg at wic-ita: Exploring the relation between semantic shifts and equivalences in translation, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[24] F. Periti, H. Dubossarsky, The time-embedding travelers@wic-ita, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

The Time-Embedding Travelers at WiC-ITA

Francesco Periti[1], Haim Dubossarsky[2]

[1]University of Milan, Italy
[2]Queen Mary University of London, England

Abstract

The WiC-ITA shared task aims to determine whether a word appearing in two distinct sentences carries the same meaning. The task consists of two subtasks: binary classification (Subtask 1) and ranking (Subtask 2). Each subtask is designed in both a monolingual (Italian) and multilingual (Italian-English) setting. In this report, we present the results of our participation in WiC-ITA. In our experiments, we leverage the condition number of the cosine similarity matrix between XLM-R embeddings and demonstrate competitive performance, ranking among the top positions in both the monolingual and cross-lingual setting. Our results indicate that semantic information is present not only in the last layers but also across the middle layers of XLM-R and throughout the entire architecture. This suggests potential avenues for future research to explore the use of the complete set of embeddings, rather than solely relying on the embeddings extracted from the last layer(s).

Keywords

Word-in-Context, Contextualized Embeddings, Condition Number

1. Introduction

In the last decade, the use of Word Embedding techniques has improved the modeling of lexical semantics. Initially, static embedding models have been employed to encode the dominant semantics of a word into a single vector representation, i.e., word embedding (Mikolov et al., 2013 [1]). However, understanding the meaning of words in their specific contexts is a crucial task for modeling language effectively. This motivated the recent efforts to create contextualized models capable of generating different vector representations according to the context in which the words occur (Devlin et al., 2019 [2]).

Despite the growing popularity of contextualized embeddings in research fields such as Word Sense Disambiguation or Lexical Semantic Shift Detection (Scarlini et al., 2020 [3]; Montanelli and Periti, 2023 [4]), Word-in-Context (WiC) benchmarks that specifically focus on the dynamic of word semantics are relatively recent. The first WiC benchmarks were limited to English (Pilehvar et al., 2019 [5]; Loureiro et al., 2022 [6]). Their success prompted the development of new WiC benchmarks to cover a wider scope of languages (Raganato et al., 2020 [7]; Liu et al., 2021 [8]), test the transfer learning ability in cross-lingual settings (Martelli et al., 2021 [9]), and evaluate graded word similarity in context (Armendariz et al., 2020 [10]).

The WiC-ITA shared task at EVALITA 2023 provides a novel benchmark for evaluating WiC for both a monolingual (L) setting in Italian and a cross-lingual (XL) setting from Italian to English (Cassotti et al., 2023 [11]; Lai et al., 2023 [12]). Inspired by the previous work, WiC-ITA challenges its participants with two sub-tasks:

1. Binary Classification: to establish if a target word w occurring in a pair of sentences $\langle s_1, s_2 \rangle$ has the same meaning or not (Subtask 1);
2. Ranking: to rank the pair of sentences $\langle s_1, s_2 \rangle$ by the degree of similarity of the target word's meaning (Subtask 2).

We participated in both Subtask 1 and 2 as the The Time-Embedding Travelers, alongside three other participants and one baseline system. Each participant was allowed to make three different submissions. In our experiments, we investigated the potential of multilingual pre-trained models in both the L- and XL-WiC setting. Given a pair of sentences $\langle s_1, s_2 \rangle$ and a target word w, our submitted systems compare the word semantics by using the cosine similarity matrix between the XLM-R word embeddings of w extracted from different layers. In particular, we use the condition number of the cosine similarity matrix to assess the degree of semantic similarity between two instances of the word.

In the official ranking, our evaluation phase submission ranked 2nd for the L-Subtask1, 1st for the XL-Subtask1, 2nd for the L-Subtask2, and 1st for the XL-Subtask2. In this paper, we extensively evaluate the effectiveness of our systems on two different multilingual models, namely multilingual BERT (mBERT) and XLM-RoBERTa (XLM-R). Our code is available at https://github.com/FrancescoPeriti/WiC-ITA.

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
francesco.periti@unimi.it (F. Periti); h.dubossarsky@qmul.ac.uk (H. Dubossarsky)
0000-0001-8388-2317 (F. Periti); 0000-0002-2818-6113 (H. Dubossarsky)
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper47.pdf

2. Background and motivation

BERT is a powerful contextualized model that leverages the Transformer encoder to capture the contextual semantics of words (Devlin et al., 2019 [2]; Vaswani et al., 2017 [13]). Typically, the success of BERT is attributed to its multi-layer (e.g., 12) and multi-head (e.g., 12) self-attention blocks. However, most of the SOTA work only uses the outputs of the final layer(s) (i.e., word embeddings) as input for solving NLP tasks, while ignoring the output of the earlier layers. As a result, the role of different embedding layers for representing the semantics of word occurrences is still unclear. Recently, a limited number of studies have been conducted to explore the nature and characteristics of the BERT embeddings. In particular,Jawahar et al. (2019) [14] indicate that BERT's lower layers capture surface features pertaining to phrase-level information, middle layers capture syntactic features, and higher layers capture semantic features. Devlin et al. (2019) [2] report that combining the last four hidden layers could be beneficial for mainstream tasks such as Named Entitiy Recognition. Ethayarajh (2019) [15] demonstrates that the geometry of the embedding space exhibits anisotropy, meaning that the embeddings of all layers occupy a narrow cone within the vector space. Other work involves probing tasks, as proposed in Hewitt et al. (2019) [16]. These tasks consist of training an auxiliary classifier on top of a model, where the contextualized embeddings serve as features to predict syntactic (e.g., part-of-speech tags) and semantic (e.g., word relations) properties of words. The idea is that if the auxiliary classifier accurately predicts a linguistic property, we can assume that the property is encoded in the tested model. In line with this work, Coenen et al. (2019) [17] investigate the capability of word sense prediction and indicate that earlier-layer embeddings contain significantly more semantic information than conventionally believed.

Thus, our experiments are motivated by the latter finding and inspired by linguistic research that highlights the influential role of morphology and syntax in shaping word meanings (Wysocki and Jenkins, 1987 [18]). In this paper, we challenge the hypothesis that word meanings should be investigated by considering the full output of pre-trained models to encompass not only semantic features of the last layers but also the intricate interplay of semantic, surface, and syntactic features present in the middle and lower layers of the contextualized models.

3. System overview

Our system is a simple threshold-based classifier based on the similarity of two sets of word vectors. In particular, given a pair of sentences $\langle s_1, s_2 \rangle$, and a target word w, we use the output embeddings of a contextualized embedding model to compute a continuous similarity score. This score indicates the extent to which the target w carries the same meaning in the sentences s_1 and s_2.

More precisely, consider a sentence s that contains the word w. Given a contextualized model M, a vector representation of w is extracted from every layer of the model M. This way, the word w in the sentence s is associated with a set of contextualized embeddings denoted by H. It's worth noting that $H \in \mathbb{R}^{n \times d}$, where n is the number of encoders of the model M (e.g., 12) and d is the dimension of the embeddings (e.g., 768). As a result, we denote as H_1 and H_2 the contextualized embeddings of w extracted from the sentences s_1 and s_2, respectively.

In order to evaluate the similarity of the word w in the contexts s_1 and s_2, we collect the pairwise cosine similarities between H_1 and H_2. We denote as S the similarity matrix between H_1 and H_2 (see Figure 3 as an example [1]) . Our hypothesis is that taking into account information from all layers at once will provide a richer and more comprehensive rapport of the nature of usage similarities of a word between the two sentences. We hypothesize that because many layers are known to capture relevant semantic information, we should consider as many of them as possible together, as they may contain more comprehensive information than a single layer comparison approach.

In order to tap into this pool of similarity scores encoded within S (that contains 144 times more information than a single layer) we use a measure called the *condition number*. The condition number of a matrix, which was already successfully applied in other domains in NLP (Dubossarsky et al., 2020 [19]), provides us with a unified measure that takes into account the many similarities scores between the representations of w of the pair s_1 and s_2 throughout the different layers.

Originally, the condition number of a matrix was used to measure its sensitivity to perturbations, or small changes, in its input. A large condition number indicates that the matrix is ill-conditioned, meaning it is sensitive to small perturbations. On the other hand, a small condition number indicates that the matrix is well-conditioned, meaning that small changes will not affect it much.

In the setting of the WiC task, we interpret the condition number of a similarity matrix as associated with the stability of meaning between the two sentences. Higher similarity scores in S overall indicate two similar word usage and are expected to produce lower (and better) condition number. On the other hand, less similar and more

[1] Corresponding record.
Italian: E siccome mi lascia gps, ecoscandaglio, tutta l'attrezzatura [...] è un *affare*. Fatto. / La rivista nordamericana segnala come presunti sospetti [...] per gli *Affari* Latinoamericani.
English: And since he leaves me gps, depth sounder, all the equipment [...] it's a *bargain*. Done. / The North American magazine lists as alleged suspects [...] for Latin American *affairs*.

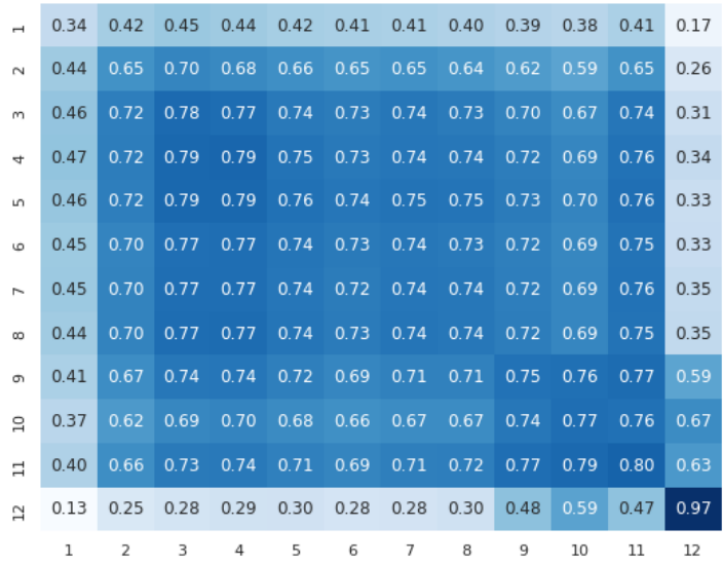


Figure 1: Pairwise similarity matrix (rounded to two decimal places) between the 12 XLM-R embeddings of the target word w in an arbitrary pair $\langle s_1, s_2 \rangle$ (the two original sentences appear in footnote 1 above).

varied similarity scores indicate more unrelated usages resulting in a higher (and worse) condition number.

The condition number of a matrix is defined as the multiplication of the matrix's norm by the norm of its reciprocal (i.e., the inverse of the matrix). The norm could be Euclidean norm, Max norm, Frobenius norm, etc. In our experiments, we calculate the condition number (COND) of the similarity matrix S using the Frobenius norm as follows:

$$\text{COND}\ (S) = \|S\|_F \cdot \|S^{-1}\|_F$$

When we compute the condition number from the similarity matrix S, we assess the degree of semantic similarity sim of a word w in each pair $\langle s_1, s_2 \rangle$ as $sim = \text{COND}(S)$ [2].

Furthermore, we also investigate the similarity sim by considering only a subset of S. We test COND_F, COND_M, and COND_L based on the similarities collected from the first, middle, and last four layers of the model M, respectively.

For the sake of comparison, we set as reference baselines the cosine similarity (CS) of the w embeddings extracted from all the layers of the model M individually, meaning that we compute n different CS scores as

$$\text{CS}_i(H_1[i], H_2[i]) = \frac{H_1[i] \cdot H_2[i]}{\|H_1[i]\|\|H_2[i]\|},$$

with $i \in 1, ..., n$. Additionally, we compute the cosine similarity CS_{AVG} between the word embeddings obtained by averaging the last four embeddings of H_1 and H_2, respectively.

In line with the WiC-ITA guidelines, we compute the Spearman correlation between the estimated similarity scores and the gold answers. This serves as the evaluation metric for Subtask 2.

In Subtask1, our binary predictions are derived from the similarity scores obtained in Subtask 2. We employ a threshold-based classifier, selecting the threshold value that optimizes the F1 score on the set of sentence pairs used as training set.

4. Experimental setup

In this task, we compared two different contextualized multilingual models, namely mBERT (Devlin et al., 2019 [2]), and XLM-R (Conneau et al., 2020 [20]). We use the Transformers library by HuggingFace to extract contextual word embeddings from mBERT and XLM-R models without performing any fine-tuning stage (Wolf et al., 2020 [21]). We use the base versions, with 12 layers and 768 hidden dimensions: *bert-base-multilingual-cased*, and *xlm-roberta-base*, respectively.

Given a target word w and a pair $\langle s_1, s_2 \rangle$. The acquisition of contextual embeddings is done by feeding the models with the sentences s_1 and s_2 individually. For every sentence, we extract the token embedding for the target word w from each layer of the model. Due to the

[2]For ease of interpretation, in our experiment, we utilized the -COND metric. We chose to associate smaller condition numbers with unrelated usages (annotated as 1), while larger numbers with identical usages (annotated as 4).

	Spearman		**Precision**		**Recall**		**F1 score**		**Threshold**
Measures	*Train*	*Test*	*Train*	*Test*	*Train*	*Test*	*Train*	*Test*	*Train*
COND	**0.520**	**0.519**	0.741	0.734	0.742	0.735	0.741	0.734	-6973.703
COND_M	0.514	0.513	**0.747**	**0.743**	**0.746**	**0.742**	**0.747**	**0.742**	-1195.522
COND_L	0.493	0.493	0.733	0.730	0.736	0.732	0.734	0.730	-972.068
CS_{10}	0.408	0.406	0.707	0.701	0.721	0.715	0.708	0.702	0.860
CS_{11}	0.404	0.403	0.704	0.698	0.709	0.703	0.706	0.700	0.895
CS_{AVG}	0.397	0.395	0.701	0.694	0.714	0.707	0.702	0.694	0.897
CS_9	0.395	0.393	0.699	0.693	0.712	0.705	0.701	0.694	0.853
CS_8	0.390	0.388	0.694	0.687	0.703	0.695	0.697	0.690	0.839
CS_7	0.370	0.367	0.695	0.690	0.698	0.693	0.696	0.691	0.839
CS_6	0.369	0.366	0.684	0.676	0.691	0.682	0.686	0.678	0.819
CS_{12}	0.342	0.339	0.683	0.679	0.701	0.697	0.686	0.682	0.987
CS_5	0.325	0.322	0.666	0.661	0.680	0.675	0.670	0.665	0.816
CS_4	0.254	0.252	0.633	0.630	0.657	0.654	0.639	0.636	0.799
COND_F	0.226	0.228	0.638	0.634	0.652	0.648	0.643	0.638	-548.832
CS_3	0.193	0.191	0.615	0.607	0.652	0.644	0.621	0.614	0.781
CS_2	0.156	0.153	0.606	0.597	0.639	0.630	0.613	0.605	0.694
CS_1	0.125	0.122	0.605	0.597	0.627	0.618	0.612	0.604	0.503

Table 1
Cross-validation analysis: average scores of WiC-ITA evaluation metrics across 100 different train-test splits. We report in bold the best result for each metric and data set.

byte-pair input encoding scheme employed by BERT-like models, some tokens may not correspond to complete words but rather to word pieces. In such cases, when a word is split into multiple tokens, we build a single word embedding by averaging the embeddings of its constituent word pieces.

Finally, to assess the graded word similarity in the context of a pair of sentences, we calculate similarity scores between the contextualized embeddings of the target word under consideration (See Section 3).

5. Experimental results

In our submissions, we rely on XLM-R as it proved to be more effective than mBERT. To maximize the performance of our system, we leverage the available *train* and *dev* set as a whole. In particular, we randomly generate 100 different train-test splits, with sizes of 2000 and 1305 respectively (equivalent to 60% and 40% of the full dataset). We conduct cross-validation on these 100 splits to validate the use of COND for Subtask2. Additionally, we leverage cross-validation to determine the optimal threshold for Subtask1, meaning that we rely on the average of the 100 best thresholds obtained during cross-validation. The average scores of Spearman correlation, Precision, Recall, and F1 score are presented in Table 1 for each tested measure. For Subtask1 and 2 and for both the L and XL setting, our three submissions correspond to the top three measures based on the F1 score and Spearman correlation, respectively (i.e., COND, COND_M, COND_L).

For the sake of comparison, Table 2 presents the preliminary performance achieved during the development phase with both XLM-R and mBERT over the Dev and Train sets.

Motivated by the superior results achieved during the development phase, we relied on XLM-R for our final submissions. In particular, we submit the predictions obtained with COND, COND_M, COND_H. However, in Table 2, it is worth noting that COND also emerged as the leading measure for the mBERT model, proving its consistency. Moreover, we note that for the WiC-ITA task, the embeddings from the last layer of both XLM-R and BERT, as well the embeddings derived by the aggregation of the last four layers, are not as effective as those from other layers. For instance, it is interesting to observe that layer 8 seems to be effective for Subtask1.

In the final evaluation leaderboard for the WiC-ITA task, we ranked 2$^{\text{nd}}$ for L-Subtask1, 1$^{\text{st}}$ for XL-Subtask1, 2$^{\text{nd}}$ for L-Subtask2, and 1$^{\text{st}}$ for XL-Subtask2. The leaderboard is reported in Table 3.

Our final results at WiC-ITa demonstrate that COND effectively captures semantic features of word meanings and can be successfully applied to tasks like WiC. Based on our development results, we assert that COND consistently outperforms the CS measure computed over individual contextualized embeddings, for Subtask 1 and 2 in both in L and XL setting. This is particularly interesting considering that CS is commonly utilized in NLP tasks to capture contextual semantics in contextualized embeddings.

Finally, COND_M consistently achieves good results by considering medium layers alone. These results are in line with the findings of Coenen et al. (2019) [17], and suggest that the middle layers of BERT-like models contain valuable information for effectively representing meaning. Therefore, future work should explore the application of COND for WiC and other related NLP tasks such as Lexical Semantic Change Detection (Montanelli

	Development Phase															
	Spearman				**Precision**				**Recall**				**F1 score**			
	XLM-R		mBERT		XLM-R		mBERT		XLM-R		mBERT		XLM-R		mBERT	
Measures	*Dev*	*Train*	*Dev*	*Train*	*Dev*	*Train*	*Dev*	*Train*	*Dev*	*Train*	*Dev*	*Train*	*Dev*	*Train*	*Dev*	*Train*
COND	**0.557**	**0.509**	0.413	**0.406**	0.719	0.748	0.663	0.702	0.698	0.745	0.642	0.702	0.691	0.747	0.630	**0.702**
COND_M	0.543	0.501	0.365	0.390	**0.725**	**0.759**	0.635	0.696	**0.710**	**0.752**	0.618	0.702	**0.705**	**0.755**	0.606	0.699
COND_L	0.502	0.489	0.275	0.241	0.712	0.744	0.619	0.641	0.688	0.741	0.584	0.663	0.679	0.743	0.551	0.650
CS_{10}	0.459	0.394	0.392	0.340	0.660	0.716	0.666	0.690	0.626	0.730	0.624	0.706	0.605	0.721	0.599	0.695
CS_{11}	0.439	0.392	0.389	0.331	0.654	0.716	0.666	0.688	0.636	0.716	0.636	0.695	0.625	0.716	0.619	0.691
CS_{AVG}	0.447	0.383	0.414	0.360	0.656	0.713	0.674	0.704	0.626	0.726	0.660	0.692	0.607	0.717	0.653	0.697
CS_9	0.439	0.382	0.420	0.366	0.666	0.710	**0.682**	0.708	0.636	0.720	0.662	0.699	0.619	0.714	0.652	0.703
CS_8	0.435	0.378	**0.429**	0.385	0.667	0.708	0.679	**0.714**	0.640	0.710	**0.664**	**0.704**	0.625	0.709	**0.657**	0.708
CS_6	0.425	0.355	0.408	0.337	0.656	0.696	0.679	0.687	0.632	0.697	0.648	0.694	0.617	0.696	0.632	0.691
CS_7	0.432	0.354	0.424	0.365	0.686	0.707	0.675	0.702	0.664	0.702	0.658	0.696	0.654	0.704	0.650	0.699
CS_{12}	0.387	0.329	0.405	0.353	0.642	0.696	0.670	0.680	0.614	0.710	0.648	0.683	0.594	0.701	0.636	0.682
CS_5	0.383	0.312	0.377	0.304	0.643	0.681	0.652	0.674	0.612	0.691	0.622	0.683	0.590	0.685	0.602	0.678
CS_4	0.324	0.241	0.329	0.256	0.613	0.648	0.621	0.640	0.578	0.667	0.586	0.661	0.543	0.656	0.553	0.648
COND_F	0.180	0.233	0.286	0.245	0.591	0.654	0.617	0.669	0.570	0.661	0.598	0.678	0.544	0.658	0.581	0.673
CS_3	0.266	0.182	0.281	0.224	0.602	0.629	0.606	0.631	0.562	0.664	0.578	0.650	0.514	0.640	0.548	0.639
CS_2	0.188	0.152	0.212	0.183	0.578	0.618	0.582	0.631	0.552	0.650	0.558	0.653	0.511	0.629	0.523	0.640
CS_1	0.138	0.125	0.157	0.166	0.555	0.625	0.591	0.627	0.540	0.636	0.564	0.654	0.506	0.630	0.529	0.637

Table 2
Preliminary performance achieved during the development phase with both XLM-R and mBERT over the Dev and Train sets. We report in bold the best result for each metric, model, and data set.

		Evaluation Phase			
		Subtask1		Subtask2	
Teams	Run	*L-WiC*	*XL-WiC*	*L-WiC*	*XL-WiC*
BERT 4EVER	run1	0.530	0.490	0.340	0.160
BERT 4EVER	run2	0.560	0.520	0.300	0.150
BERT 4EVER	run3	0.560	0.490	-	-
LG	LG	**0.730**	-	0.490	-
The Time-Embedding Travelers	COND_M	0.660	0.720	0.520	**0.550**
The Time-Embedding Travelers	COND_L	0.620	**0.740**	0.490	0.530
The Time-Embedding Travelers	COND	0.670	0.730	0.550	0.540
extremITA	camoscio lora	0.510	0.540	-	-
extremITA	it5	0.610	0.620	-	-
Baseline	-	0.590	0.560	**0.570**	0.410

Table 3
Final evaluation leaderboard for the WiC-ITA task. The best results are highlighted in bold. Our rankings are as follows: 2nd for L-Subtask1, 1st for XL-Subtask1, 2nd for L-Subtask2, and 1st for XL-Subtask2.

and Periti, 2023 [4]; Tahmasebi et al., 2023 [22]).

Additionally, we observe that the lower layers of XLM-R struggle to distinguish between different meanings when considered individually. However, when incorporated into the standard COND framework, their inclusion leads to improved results compared to COND_M. This suggests that our original intuition was correct and that the embedding layers should be considered as a whole, rather than solely focusing on the last layers.

6. Conclusion

Our experiments for the WiC-ITA shared task ranked 2nd for L-Subtask1, 1st for XL-Subtask1, 2nd for L-Subtask2, and 1st for XL-Subtask2. In our submissions, we use the condition number of the cosine similarity matrix between XLM-R embeddings extracted from different layers. Our results support our initial hypothesis that leveraging all the information provided by the pre-trained model can significantly enhance performance on mainstream tasks such as WiC. Specifically, our research suggests that the embedding layers should be considered as a whole, rather than solely focusing on the last layers of the model, which is the conventional practice. This motivates further exploration of alternative measures that can effectively represent word meanings by considering sets of embeddings as input, rather than just individual vectors. Thus, we plan to integrate the condition number into our recently developed approach for Semantic Change Detection [23].

Acknowledgments

This work has in part been funded by the project Towards Computational Lexical Semantic Change Detection supported by the Swedish Research Council (2019–2022; contract 2018-01184), and in part by the research program Change is Key! supported by Riksbankens Jubileumsfond (under reference number M21-0021).

References

[1] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, in: Y. Bengio, Y. LeCun (Eds.), Proc. of ICLR, Scottsdale, Arizona, 2013.

[2] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proc. of NAACL-HLT, ACL, Minneapolis, Minnesota, 2019, pp. 4171–4186.

[3] B. Scarlini, T. Pasini, R. Navigli, With More Contexts Comes Better Performance: Contextualized Sense Embeddings for All-Round Word Sense Disambiguation, in: Proc. of EMNLP, ACL, Online, 2020, pp. 3528–3539.

[4] S. Montanelli, F. Periti, A Survey on Contextualised Semantic Shift Detection, 2023. arXiv:2304.01666.

[5] M. T. Pilehvar, J. Camacho-Collados, WiC: the Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations, in: Proc. of NAACL-HLT, ACL, Minneapolis, Minnesota, 2019, pp. 1267–1273.

[6] D. Loureiro, A. D'Souza, A. N. Muhajab, I. A. White, G. Wong, L. Espinosa-Anke, L. Neves, F. Barbieri, J. Camacho-Collados, TempoWiC: An Evaluation Benchmark for Detecting Meaning Shift in Social Media, in: Proc. of COLING, ICCL, Gyeongju, Republic of Korea, 2022, pp. 3353–3359.

[7] A. Raganato, T. Pasini, J. Camacho-Collados, M. T. Pilehvar, XL-WiC: A Multilingual Benchmark for Evaluating Semantic Contextualization, in: Proc. of EMNLP, ACL, Online, 2020, pp. 7193–7206.

[8] Q. Liu, E. M. Ponti, D. McCarthy, I. Vulić, A. Korhonen, AM2iCo: Evaluating Word Meaning in Context across Low-Resource Languages with Adversarial Examples, in: Proc. of EMNLP, ACL, Punta Cana, Dominican Republic, 2021, pp. 7151–7162.

[9] F. Martelli, N. Kalach, G. Tola, R. Navigli, SemEval-2021 Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation (MCL-WiC), in: Proc. of SemEval, ACL, Online, 2021, pp. 24–36.

[10] C. S. Armendariz, M. Purver, M. Ulčar, S. Pollak, N. Ljubešić, M. Granroth-Wilding, CoSimLex: A Resource for Evaluating Graded Word Similarity in Context, in: Proc. of LREC, ELRA, Marseille, France, 2020, pp. 5878–5886.

[11] P. Cassotti, L. Siciliani, L. Passaro, M. Gatto, P. Basile, WiC-ITA at EVALITA2023: Overview of the EVALITA2023 Word-in-Context for ITAlian Task, in: Proc. of EVALITA, CEUR.org, Parma, Italy, 2023.

[12] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, EVALITA 2023: Overview of the 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, in: Proc. of EVALITA, CEUR.org, Parma, Italy, 2023.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is All you Need, in: Proc. of NeurIPS, volume 30, CAI, 2017.

[14] G. Jawahar, B. Sagot, D. Seddah, What Does BERT Learn about the Structure of Language?, in: Proc of ACL, ACL, Florence, Italy, 2019, pp. 3651–3657.

[15] K. Ethayarajh, How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings, in: Proc. of EMNLP-IJCNLP, ACL, Hong Kong, China, 2019, pp. 55–65.

[16] J. Hewitt, P. Liang, Designing and Interpreting Probes with Control Tasks, in: Proc. of EMNLP-IJCNLP, ACL, Hong Kong, China, 2019, pp. 2733–2743.

[17] A. Coenen, E. Reif, A. Yuan, B. Kim, A. Pearce, F. Viégas, M. Wattenberg, Visualizing and Measuring the Geometry of BERT, Curran Associates Inc., Red Hook, NY, USA, 2019.

[18] K. Wysocki, J. R. Jenkins, Deriving word meanings through morphological generalization, Reading Research Quarterly (1987) 66–81.

[19] H. Dubossarsky, I. Vulić, R. Reichart, A. Korhonen, The Secret is in the Spectra: Predicting Cross-lingual Task Performance with Spectral Similarity Measures, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 2377–2390.

[20] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised Cross-lingual Representation Learning at Scale, 2020. arXiv:1911.02116.

[21] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-Art Natural Language Processing, in: Proc. of EMNLP, ACL, Online, 2020, pp. 38–45.

[22] N. Tahmasebi, H. Dubossarsky, Computational modeling of semantic change, 2023. arXiv:2304.06337.

[23] F. Periti, A. Ferrara, S. Montanelli, M. Ruskov, What is Done is Done: an Incremental Approach to Semantic Shift Detection, in: Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 33–43. URL: https://aclanthology.org/2022.lchange-1.4. doi:10.18653/v1/2022.lchange-1.4.

LG at WiC-ITA: Exploring the relation between semantic distance and equivalence in translation.

Lorenzo Gregori

University of Florence

Abstract

The Word in Context task has been addressed here on the basis of a simple intuition: if the same lemma has different senses in two different contexts, it tends to be translated (in other languages) with two different lemmas; conversely, in the case of the same sense, translation lemmas tend to be the same. The proposed methodology is based on the translation of sentence pairs in 21 languages, and the use of a SVM classifier/regressor. Obtained results are excellent in binary classification and average in regression tasks.

Keywords

EVALITA, Word In Context, Machine Translation, Semantics

1. Introduction

The Word-in-Context (WiC) task is a new task [1, 2], that aims to identify if the same word used in two contexts has the same sense in both contexts or if it's used with different senses. Interestingly, in the current EVALITA task [3, 4] sentences are manually judged by several annotators.

The method chosen to solve this task is completely different from the previously used methods, mostly based on the use of pre-trained language models: see recent works by Alan Ansell and colleagues, and Qianchu Liu and colleagues, among others [5, 6]. Actually, in the proposed approach language models are used not to directly accomplish the given task, but to provide accurate sentence translations in several languages.

The intuition behind the proposed methodology is that a word with the same sense in two contexts is probably translated with the same lemma in a target language; otherwise, a word that is used in two different senses tends to be translated with two different lemmas. This idea has been first explored by Gale, Church, and Yarowsky in the early 1990s [7, 8], and is a basic concept behind the work on semantic spaces produced by Melissa Bowerman and colleagues in the early 2000s [9, 10], and the IMAGACT Ontology of Action [11, 12].

Gale and colleagues showed that having a text translated into another language can be useful for word sense disambiguation, given that a word with two different senses is frequently translated with two different words in the target language. Bowerman and colleagues analyzed the variation of event categorization in different languages, by studying the semantic variation of general verbs "cut" and "break". Their work showed how languages use their own verbs to partition differently the semantic space related to these two events. IMAGACT[1] highlights the relation between different semantic types of an action verb, and different groups of verbs usable to predicate them. From this resource, we can clearly observe that, in general, the verb set allowed for one type is not the same set allowed for another type, even if the two types belong to the same verb. This is a shared property across languages: it occurs in Italian, English, and in many other languages.

The proposed approach is based on high-quality translation of the provided sentences in several languages; translated sentences are then word-aligned to the original ones, and lemmatized. With this data, it's easy to verify if the lemma used to translate the target word in the two original sentences is the same or not. This binary information repeated for each language is used to compile a feature vector. Then, a Support Vector Machines (SVM) classifier and an SVM regressor are trained on these vectors to decide if (or how much) the two word senses are different.

2. Word semantics in translation

Translating a word to another language is not a trivial task, because it's pretty rare to find a word in the target language with exactly the same meaning as the original word. More likely, there are several possible translators, each one suitable for some contexts and not for others. Moreover, even at the lower level of word senses, it is hard to find perfect matches between two languages, given that languages partition semantic spaces in their own way [13, 14].

[1] http://www.imagact.it/

EVALITA'23: Evaluation of NLP and Speech Tools for Italian, September 7-8, 2023, Parma, Italy

lorenzo.gregori@unifi.it (L. Gregori)

0000-0001-9208-2311 (L. Gregori)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper48.pdf

	$s(w_a) = s(w_b)$	$s(w_a) \neq s(w_b)$
$s(t_1) = s(t_2)$	$p =?$	$p =?$
$s(t_1) \neq s(t_2)$	$p =?$	$p =?$

Table 1
The set of 21 translation languages used to solve the WiC task.

These premises, plus the fact that the same word can have different meanings, and different words can be synonyms, make the picture of semantics in translation very complex.

Consider two occurrences of the same word (in different contexts), w_a and w_b, translated to another language with two words, t_1 and t_2. Then, w_a and w_b can have the same sense or different senses; t_1 and t_2 can be the same word or different words, and in both cases, they can have the same sense or two different senses (they can be synonyms if they are different word with the same sense, or polysemous if they are the same word with different senses).

So, the 4 cases represented in Table 1 ($s(x)$ means the sense of the word x) are all possible, both if t_1 and t_2 are a single word, or if they are different words. It is the probability of these cases that is undetermined ($p =?$).

The assumption behind this experiment is that the following two cases are more frequent than others:

- $(s(w_a) = s(w_b)) \wedge (s(t_1) = s(t_2))$ if t_1 and t_2 are the same word;
- $(s(w_a) \neq s(w_b)) \wedge (s(t_1) \neq s(t_2))$ if t_1 and t_2 are two different words.

If this is true for most sentences in most languages, then the use of many translation languages provides reliable information to identify semantic distance between w_a and w_b.

The two examples reported below are derived from the test set and aim to clarify the idea behind the proposed methods, and the practical issues.

Example 1. The target word has two different senses. The following two sentences have been labeled as 0 (i.e. target word belonging to different senses):

1. *Chi ha intenzioni meno serie, troverà godibile il tour fotografico del **complesso** , completato dalla visita virtuale del museo...*
2. *...la Camera dei deputati aveva approvato un **complesso** di disposizioni leggermente diverse da quelle recepite dalla n. 180.*

In fact, in most languages two different lemmas have been used in translation:

- French uses **complexe** in (1) and **ensemble** in (2);
- Finnish uses **kompleksi** in (1) and **joukko** in (2);
- English uses **complex** in (1) and **set** in (2).

The adoption of two different lemmas to translate these occurrences of "complesso" is spread over several languages, but, of course, there are some exceptions. In Bulgarian, for example, in both of the sentences, "complesso" is translated with a unique word "комплекс", similar to Italian.

Dealing with translations, it's not possible to rely on a perfect word-to-word alignment: it can occur that multiple words are translated with only one word in the target language, or, vice-versa, a unique word translated with more words, or some words are just omitted in translation. Moreover, some alignment errors must be expected by the word aligner. In this example, the word in (1) is translated into Lithuanian with "kompleksas", while in (2) there is not any word aligned to "complesso". It could be due to the linguistic properties of Lithuanian or to an error in the word alignment.

Example 2. The target word has the same sense. The following two sentences have been labeled as 1 (i.e. target word belonging to the same sense):

1. *inserendo in questa **maschera** la parola greca "mache" (battaglia) si otterranno tutti i termini collegati...*
2. *Questa **maschera** consente di visualizzare alcune informazioni in forma sintetica: il titolo del documento,...*

As in the previous example, most languages used the same lemma to translate "maschera" in these two contexts:

- Lithuanian used **kaukė** for both (1) and (2);
- Malay used **topeng** for both (1) and (2);
- Spanish used **máscara** for both (1) and (2).

Icelandic used two different words: **sjónarhóll** for (1), and **gríma** for (2).

3. Description of the system

The system proposed to solve the WiC-ITA task is divided into two parts: (a) the creation of a feature vector related to each sentence pair, and (b) the training of a machine learning algorithm on the feature vectors.

3.1. Feature vectors

Given two sentences $s1$ and $s2$, containing both an occurrence of the same lemma; these occurrences are the target words $w1$ and $w2$. For each of the 21 languages (l) considered (see Table 2 for the full list), the algorithm performs the following steps:

Albanian	French	Lithuanian
Armenian	German	Macedonian
Bulgarian	Greek	Malay
Czech	Hindi	Manx
Danish	Hungarian	Russian
English	Icelandic	Spanish
Finnish	German	Macedonian

Table 2
The set of 21 translation languages used to solve the WiC task.

1. Translate $s1$ in language $l \rightarrow t1$;
2. Translate $s2$ in language $l \rightarrow t2$;
3. Align at word-level $t1$ to $s1 \rightarrow t1a$;
4. Align at word-level $t2$ to $s2 \rightarrow t2a$;
5. Lemmatize $t1a$ and find the lemma of $t1a$ aligned to $w1 \rightarrow w1a$;
6. Lemmatize $t2a$ and find the lemma of $t2a$ aligned to $w2 \rightarrow w2a$;
7. Assign 0.5 if $w1a$ or $w2a$ are non-words;
   Assign 0 if $w1a = w2a$;
   Assign 1 if $w1a \neq w2a$;

The result of this algorithm is a feature vector v with a number {0, 0.5, 1} for each language. The size of v is the number of languages used (21 in the proposed experiment).

As an example, consider the following sentences pair, belonging to the test set:

- *... che vi sia alla base un accordo tra i coniugi, soprattutto in relazione all'educazione del* ***minore****.*
- *Infatti ogni religione istituzionalizzata annette importanza maggiore o* ***minore*** *alla propagazione ... dei suoi riti.*

After the algorithm execution, the v vector appears as below (only the first 10 elements are reported).

en	sv	el	id	ru	hi	es	de	hy	bg	
(1	1	1	0.5	1	1	0	1	1	1	...)

Most of the values are 1, meaning that many languages translated the word *minore* with different words; there is a 0.5 in 4th position, meaning that there is a missing alignment of the word *minore* with a corresponding Indonesian (id) word (in at least one of the two sentences); the value of 0 (7th position) means that in Spanish (es) *minore* has been translated with the same word in both the contexts.

3.2. External tools

The key point of this algorithm is the sentence translation engine, which needs to have a high accuracy, and ensure a really contextualized translation; moreover, the translation system must be available in several languages. To this aim, Opus-MT system [15, 16] has been selected: it is a state-of-the-art system of neural machine translation, for which more than 1,400 pretrained models are freely available[2]: each model is specifically trained on a language pair.

To align sentences and translations at the word level, the multilingual word aligner created by Dou & Neubig has been selected [17]: the pretrained model aligned on multilingual BERT is freely available online[3].

The lemmatizer used for this work is Simplemma[4], a tool available as a Python library that performs sentence lemmatization in over 50 languages.

Both the aligner and the lemmatizer have been chosen to be multilingual ready-to-use tools, that can be easily included in the implemented pipeline.

label	**inst.**
0	806
1	1,999

Table 3
Number of instances (sentence pairs) for each label in the training set of the binary classification task.

3.3. Training the machine

Support Vector Machines algorithms have been used on the current dataset both for regression and classification.

Algorithms have been trained on the Italian dataset and tested on the Italian dataset only.

Binary classification task After the conversion of each pair of input sentences in a vector, a classifier was trained on the training set and tested on the development set. An SVM classifier with a linear kernel was chosen for its good performance on this task. The complexity hyperparameter (C) has been tuned on the dev set, obtaining the best results with $C = 0.1$.

The dataset has been balanced with a random undersampling technique, to train the algorithm with an equal number of examples of positive and negative classes (i.e. the sentences where the target word has the same sense, and the ones where the target word has different senses). As reported in Table 3, the training set for the classification task is highly unbalanced, with a proportion between the two classes of 29% - 71%.

Ranking task

The ranking task has been solved with an SVM regressor, using a Gaussian kernel, that led to better performance

[2]https://huggingface.co/Helsinki-NLP
[3]https://github.com/neulab/awesome-align
[4]https://pypi.org/project/simplemma/

score	inst.	score	inst.
1	387	3	303
1.5	300	3.5	723
2	119	4	973

Table 4
Number of instances (sentence pairs) for each score in the training set of the ranking task.

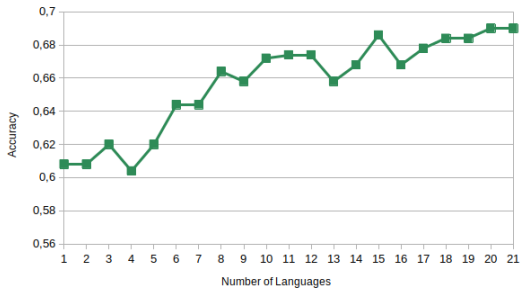


Figure 1: Accuracy growth of the binary classification, when increasing the number of languages.

than the linear kernel[5].

A preliminary analysis of the training set for the ranking task highlighted that data are highly unbalanced; moreover, the score values assigned to each pair of sentences are exactly 6: {1, 1.5, 2, 3, 3.5, 4 } (see Table 4).

It is possible to identify two sources of biases: (a) 71% of instances have a high score (3 to 4), while only 29% has a low score (1 to 2); there is an increasing number of scores, moving form central scores (2 and 3) to the extreme scores (1 and 4). This suggested performing a random undersampling to balance the sentences assigned to each score and train the algorithm in a reduced unbiased space.

4. Results

The first issue that emerged in the training was about the number of languages that should be used to obtain higher accuracy. So, at first, the algorithm has been tested using an increasing number of languages (1 to 21).

The graph reported in Fig. 1 shows the change in accuracy of the binary classification algorithm with respect to the number of languages used for training. This graph is based on the development set.

In general, we can see that with just one language the algorithm accuracy is about 0.60, moving up towards 0.70 as the number of languages increases. The choice of 21 languages seems reasonable, considering the trend of the curve, which starts flattening with more than 10 languages.

[5]Complexity hyper-parameter C is tuned to 0.1, as in the classification problem

Table 5
Best results of the Italian WiC task.

Team	Binary	Ranking
BERT4EVER	0.56	0.34
LG	**0.73**	0.49
The Time-Emb. Travelers	0.67	0.55
extremITA	0.61	-
baseline	0.59	**0.57**

Results of the Italian task are resumed in Table 5, where the highest accuracy for each team is reported. The results obtained in the classification task (on the test set) is an accuracy of 0.73, which is very high for a WiC competition [2]. Conversely, this algorithm didn't emerge in the ranking task, obtaining an average score of 0.49, which is below the baseline.

5. Discussion

The high accuracy obtained in the classification task with only features related to lemma equivalences in translation is, first of all, a piece of strong evidence to support some linguistic theory about semantic spaces.

The implemented model is easy to interpret, and quick to train. This makes the proposed system, a very flexible tool to perform other experiments, like changing the number and the type of languages used, finding the minimal set of languages with the maximum discriminative power, and so on. It would be also interesting to try to apply on another language the model trained in one language, as suggested by the task organizers.

The proposed algorithm can be improved:

- The number of languages is probably enough to reach the maximum accuracy with feature vectors; otherwise the set of used languages could be changed, by introducing, for example, some Asian languages, like Chinese, Japanese, or Korean, that probably would bring a big contribution to this task;
- Alignment and lemmatization could be improved, by using, for each language, a state-of-the-art tool that is specifically tuned for that language; this would probably lead to results that are better than the ones obtained with multi-language tools.

About the computational cost of the proposed approach, it is completely moved from the training stage (which has almost no cost) to the feature extraction: in fact, the computationally intense stage is the neural machine translation in 21 languages. This task is also very slow, but easy to parallelize, using simultaneous translation engines. Interestingly enough, once the vectors have been compiled for the dataset in use, all the experiments

with changes in the types or the number of languages are not costly.

References

[1] M. T. Pilehvar, J. Camacho-Collados, Wic: the word-in-context dataset for evaluating context-sensitive meaning representations, arXiv preprint arXiv:1808.09121 (2018).

[2] A. Raganato, T. Pasini, J. Camacho-Collados, M. T. Pilehvar, Xl-wic: A multilingual benchmark for evaluating semantic contextualization, 2020. arXiv:2010.06478.

[3] P. Cassotti, L. Siciliani, L. Passaro, M. Gatto, P. Basile, Wic-ita at evalita2023: Overview of the evalita2023 word-in-context for italian task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[4] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[5] A. Ansell, F. Bravo-Marquez, B. Pfahringer, An elmo-inspired approach to semdeep-5's word-in-context task, in: Proceedings of the 5th Workshop on Semantic Deep Learning (SemDeep-5), 2019, pp. 21–25.

[6] Q. Liu, F. Liu, N. Collier, A. Korhonen, I. Vulić, Mirrorwic: On eliciting word-in-context representations from pretrained language models, arXiv preprint arXiv:2109.09237 (2021).

[7] W. A. Gale, K. Church, D. Yarowsky, Using bilingual materials to develop word sense disambiguation methods, in: Proceedings of the Fourth Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, 1992.

[8] W. A. Gale, K. W. Church, D. Yarowsky, A method for disambiguating word senses in a large corpus, Computers and the Humanities 26 (1992) 415–439.

[9] M. Bowerman, Why can't you "open" a nut or "break" a cooked noodle? learning covert object categories in action word meanings, in: Building object categories in developmental time, Psychology Press, 2005, pp. 227–262.

[10] A. Majid, J. S. Boster, M. Bowerman, The cross-linguistic categorization of everyday events: A study of cutting and breaking, Cognition 109 (2008) 235–250. URL: https://www.sciencedirect.com/science/article/pii/S0010027708001911. doi:https://doi.org/10.1016/j.cognition.2008.08.009.

[11] M. Moneglia, G. Gagliardi, L. Gregori, A. Panunzi, S. Paladini, A. Williams, La variazione dei verbi generali nei corpora di parlato spontaneo. l'ontologia imagact, in: Proceedings of the VIIth GSCP International Conference: Speech and Corpora, 2012, pp. 406–411.

[12] M. Moneglia, M. Monachini, O. Calabrese, A. Panunzi, F. Frontini, G. Gagliardi, I. Russo, The imagact cross-linguistic ontology of action. a new infrastructure for natural language disambiguation., in: LREC, 2012, pp. 2606–2613.

[13] M. Bowerman, S. Choi, et al., Shaping meanings for language: Universal and language-specific in the acquisition of spatial semantic categories, Language acquisition and conceptual development 3 (2001) 475–511.

[14] Y. Wilks, N. Ide, Making sense about sense. word sense disambiguation, algorithms and applications (pp. 47–73), 2007.

[15] J. Tiedemann, S. Thottingal, Opus-mt–building open translation services for the world, in: Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, European Association for Machine Translation, 2020.

[16] J. Tiedemann, M. Aulamo, D. Bakshandaeva, M. Boggia, S.-A. Grönroos, T. Nieminen, A. Raganato, Y. Scherrer, R. Vazquez, S. Virpioja, Democratizing machine translation with opus-mt, arXiv preprint arXiv:2212.01936 (2022).

[17] Z.-Y. Dou, G. Neubig, Word alignment by fine-tuning embeddings on parallel corpora, in: Conference of the European Chapter of the Association for Computational Linguistics (EACL), 2021.

DisCoTex at EVALITA 2023: Overview of the Assessing DIScourse COherence in Italian TEXts task

Dominique Brunato[1], Davide Colla[2], Felice Dell'Orletta[1], Irene Dini[1,3], Daniele Paolo Radicioni[4] and Andrea Amelio Ravelli[5]

[1]Istituto di Linguistica Computazionale "Antonio Zampolli", Pisa - ItaliaNLP Lab
[2]Dipartimento di Studi Storici - Università degli Studi di Torino
[3]University of Pisa
[4]Dipartimento di Informatica - Università degli Studi di Torino
[5]Dipartimento di Lingue, Letterature e Culture Moderne - Università di Bologna

Abstract

The Assessing DIScourse COherence in Italian TEXts (DisCoTeX) task is the first shared task focused on modelling discourse coherence for Italian real-word texts, which has been proposed for the first time at EVALITA 2023. Providing two different datasets from different textual genres, we arranged the task into two independent tasks: a more traditional one, aimed at evaluating whether models are able to distinguish well-organized documents from corrupted ones and a less explored one, which assesses the models' performance on texts evaluated for coherence by human raters. In this paper, we describe the datasets released, we discuss the different approaches tackled by the participating systems and provide a first analysis of the obtained results.

Keywords

text coherence, Italian language, computational modeling, evaluation campaign, dataset

1. Introduction and Motivation

Coherence is a key property of any well-organized text and it plays a crucial role in human discourse processing. Indeed, as individuals process unfolding text, they are required to assemble information from single sentences and to draw inferences between and among them in order to create a meaningful mental representation of the whole text. According to the tripartite model developed by [1], this is the outcome of a three-step process in which readers construct multileveled memory representations of a text, encoding different, and progressively more abstract, information at each level. From this perspective, coherence is an inherently psychological construct, thus very hard to be modelled; however, it also has a counterpart at the level of linguistic content and structure, often referred to as "cohesion", a property of a text that is conveyed by signalling linguistic devices such as reference, ellipsis, discourse connectives, argument overlap, which help readers make explicit the logical links between different units in texts.

As regards the computational modelling, coherence has been widely investigated in the Natural Language Processing (NLP) community, particularly in the "pre-deep-learning" era, with much research drawing inspiration from frameworks like the Centering Theory [2]. One popular approach in this context is the entity-grid approach, which focuses on assessing *local* coherence, specifically the transitions between adjacent sentences (see, among others, [3, 4]). More recently, also neural models have been applied to deal with both structured representations of text and unstructured text by taking advantage of neural models' ability to learn useful representations for the task, e.g. [5, 6]. Modelling coherence in natural language is of pivotal importance in a variety of downstream applications, from automatic essay scoring in language learning scenarios [7, 8], to language assessment in clinical settings [9, 10] Additionally, from the Natural Language Generation point of view, coherence is an intrinsic evaluation metric to assess the quality of generated texts. An emerging area of interest pertains to the interpretability of modern deep neural networks. In this respect, while existing work on probing pre-trained language models has largely focused on sentence-level properties, the ability of these models to encode discourse and pragmatic phenomena is still unclear [11, 12, 13].

Recognizing the fundamental role that coherence plays across a variety of scenarios and the challenges in developing a unified metric to quantify this concept, DisCoTeX, organized in the context of the 8th evaluation campaign of NLP and speech tools for the Italian language (EVALITA 2023) [14] intends to encourage research on automatic discourse coherence modeling with empha-

EVALITA 2023: 8th Evaluation Campain of NLP and Speech Tools for Italian, Semptember 07–08, 2023, Parma, Italy
dominique.brunato@ilc.cnr.it (D. Brunato); davide.colla@unito.it (D. Colla); felice.dellorletta@ilc.cnr.it (F. Dell'Orletta); irene.dini@ilc.cnr.it (I. Dini); daniele.radicioni@unito.it (D. P. Radicioni); andreaamelio.ravelli@unibo.it (A. A. Ravelli)
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper49.pdf

sis on the Italian language.

2. Definition of the Task

Drawing inspiration from existing coherence modeling literature, the DisCoTeX task was designed with the intention of addressing two distinct scenarios. The first scenario involves the evaluation of models' ability to differentiate well-structured documents from corrupted ones. The corrupted documents are typically created by either shuffling the sentence order of the original document or replacing specific linguistic elements that contribute to coherence within and across sentences, such as personal pronouns or discourse connectives. The second scenario, which has been less explored, focuses on assessing the models' performance in coherence evaluation by comparing their predictions to human raters' evaluations.

To capture these distinct scenarios, we proposed two independent sub-tasks:

- Sub-task 1 - **Last sentence classification**: This sub-task was casted as a binary classification task. Specifically, participants are presented with a prompt, which is a short paragraph consisting of approximately three consecutive sentences, and an individual sentence referred to as the target. The objective is to classify whether the target sentence, when combined with the prompt, forms a coherent or incoherent text. The negative target can either be a sentence randomly selected from a different document or a sentence extracted from the same document as the prompt, in order to introduce incremental degrees of complexity on the resolution of the task;
- Sub-task 2 - **Human score prediction**: This sub-task was framed as a regression task where participants were asked to predict the average coherence score assigned by human raters to short paragraphs. These paragraphs were evaluated in their original or artificially modified version. As shown in previous tasks on the automatic assessment of subjective phenomena [15, 16], this scenario is expected to be more challenging, as it requires modeling the human perception of coherence, which can be influenced by both linguistic and non-linguistic factors, as highlighted in previous studies [7].

For both sub-tasks, dataset were extracted from two corpora representative of two distinct domains, as described in the following section.

3. Datasets

The dataset[1] utilized for the DisCoTeX task encompasses texts sourced from two distinct origins: the Italian Wikipedia and the Italian speech transcripts section of the Multilingual TEDx corpus (mTEDx). These sources represent two different language varieties: the former is a 'standard' written variety, and the latter a 'hybrid' variety combining diverse genres (e.g., university lectures, newspaper articles, conference presentations, and TV science programs) as well as different semiotic modes, such as written, spoken, audio, and video [17]. Extensive research on genre and register variation acknowledges that written and spoken language employ distinct strategies to establish coherence within a text [18]. Therefore, we decided to evaluate systems on both these types of data.

For sub-task 1, each data sample consists of a prompt, which is a paragraph comprising three sentences, followed by a target sentence. To create the written dataset, we leveraged the existing paragraph segmentation in Wikipedia to select four-sentence paragraphs. For the spoken dataset, as mTEDx speeches lacked such internal structure, we divided all the transcripts into passages of four sentences. The target sentence is determined as the immediate continuation of the prompt, forming a coherent sample. In the case of a non-coherent passage, as previously anticipated, we selected either a sentence randomly taken from a different document or the sentence that appears ten sentences after the prompt in the same document. Each final dataset consists of 8,000 training samples and 800 test samples. Examples can be found in Table 1.

Regarding sub-task 2, the dataset construction differs slightly. In this case, for each source we extracted samples consisting solely of four-sentence paragraphs (we keep the term 'prompts' to refer to them), with half of them deliberately made incoherent through sentence perturbations. The possible perturbations, chosen with equal probability, include:

- Flip of two random sentences: each sentence of the prompt has the same probability of being flipped.
- Swap of a sentence with the next 10th of the same document from which the prompt was extracted. The first and the last sentence have double the swap probability compared to the middle two sentences to make the swapping of the first/last or a middle one equiprobable.

For the purposes of the DisCoTeX task, we selected 1,064 prompts equally balanced between the two domains.

[1]The DisCoTeX dataset is available at the following link: https://github.com/davidecolla/DisCoTex/

Prompt	Target	Class
Il regolamento del carcere era durissimo e le condizioni igieniche drammatiche. Agli ebrei erano negati i pochi diritti concessi agli altri prigionieri politici e comuni, ovvero l'ora d'aria in cortile, l'assistenza sanitaria, la possibilità di ricevere lettere e pacchi e di acquistare generi alimentari allo spaccio del carcere. Gli interrogatori degli arrestati erano condotti in uno stanzone a pian terreno, detto il "refettorio".	Qui le sevizie di ogni genere venivano inflitte soprattutto sugli ebrei che non rivelavano i recapiti o i nascondigli dei loro parenti della cui presenza a Milano o nei dintorni le SS erano venute a conoscenza tramite loro spie.	1
Ci siamo trovati a Brasilia, la capitale del Brasile; e c'erano città di tutto il mondo, dall'Australia al Giappone, all'Asia, all'Africa, agli Stati Uniti. E lì abbiamo avuto la consapevolezza che siamo un movimento che sta crescendo nel mondo e che sempre più costruisce risultati e vantaggi. Una delle più grandi città del mondo che ha fatto questa scelta è San Francisco.	Vedete in questo semplicissimo grafico, il rosso è tutto quello che prima,una decina di anni fa, andava a smaltimento.	0

Table 1
Examples extracted from the dataset of sub-task 1. The first one belongs to the Wikipedia dataset and the second one to mTEDx.

Of these, 33% (i.e. 360 prompts) are extracted from the subset of authentic prompts and 66% (i.e. 704) from perturbed ones. Examples can be found in Table 2.

As anticipated, coherence was assessed through manual annotation. Specifically, to gather human ratings of coherence, we conducted a crowdsourcing task on the Prolific[2] platform, involving Italian native speakers. Recognizing that coherence is a subjective concept influenced by the reader or listener's interpretation, we employed a gradual judgment approach, and asked them to evaluate their perception of on a Likert scale ranging from 1 to 5. The number of annotations per prompts ranged from 9 to 12, with an average of 11.75.

The resulting dataset was split into training and test samples with a proportion of 80% to 20%, respectively.

In Figure 1 we show some general statistics about collected judgments considering both the whole dataset of prompts and prompts grouped into specific sections, according to genre and perturbation. As it can be seen, prompts derived from Wikipedia texts are generally rated as more coherent by humans compared to TEDx prompts. This observation confirms previous findings with regard to the influence of genre on the perception of coherence [7]. What is particularly interesting is that this disparity is evident not only in the original form of the prompts but also in the perturbed versions. This seems to suggest that Wikipedia documents tend to exhibit a more standardized structure, including internal coherence, which remains relatively stable even with minor alterations that affect sentence order or the insertion of an intruder sentence from the same document. We plan to conduct a more in-depth analysis by examining each perturbation strategy independently to gain a deeper understanding of their individual impact on coherence.

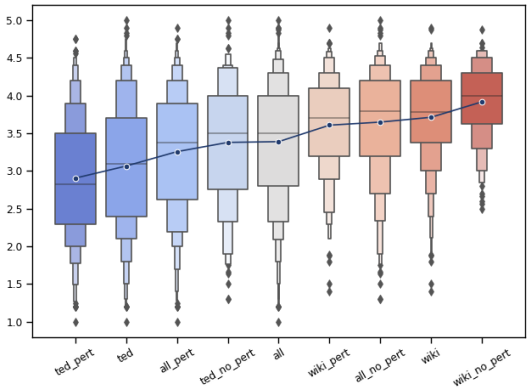


Figure 1: Overview of human judgments collected for the dataset used in sub-task 2. The plot shows the overall mean of human judgments for the whole dataset (*all*) and for respective subsets, including both coherent (**_no_pert*) and perturbed prompts (**_pert*).

3.1. Format

The DisCoTEX dataset was released as tab-separated text files. Specifically, for sub-task 1, the two data sources (i.e. Wikipedia and TED) were kept separated and, for each source, participants were provided with a file with a following structure:

- ID: a numerical identifier for the entry;
- PROMPT: textual passage made by three consecutive sentences;
- TARGET: the sentence which participants are asked to assess if it is coherent with the prompt (i.e. it is the next sentence after the prompt);

[2]https://www.prolific.co/

Text	Mean	Stdev
Le nuove idee sono una sfida che accende nel nostro cervello la stessa area che elabora le minacce fisiche. Ecco perché tendiamo a reagire con forza, a volte con aggressività, alle nuove idee. Davanti a informazioni che mettano in discussione le nostre convinzioni noi tendiamo paradossalmente a reagire rafforzandole ancora di più. Si chiamano bias cognitivi, sono molto forti e ci caschiamo tutti.	4.83	0.58
I Romani furono scommettitori appassionati, specialmente ai tempi dell'Impero Romano, e il gioco dei dadi era popolare, seppur proibito da una "Lex alearia" del 204 a.C. circa, eccetto che durante i Saturnali. Orazio derise la gioventù dell'epoca che sprecava tempo tra i pericoli del gioco invece di domare il suo cavallo e darsi alle durezze dell'inseguimento. Una di queste diceva che nessuna causa poteva essere intentata da una persona che permetteva il gioco d'azzardo nella sua casa anche se era stata imbrogliata o assalita. Le scommesse sui dadi per denaro fu l'oggetto di molte leggi Romane.	3.3	0.95

Table 2
Examples extracted from the dataset of sub-task 2. The first one is an original prompt taken from the mTEDx corpus. The second ones is a perturbed prompt from the Wikipedia corpus, with a swap between the third and the last sentence.

- CLASS: the class to be predicted (1 if the target follows the prompt, 0 otherwise).

For sub-task 2, we mixed data from the two sources and released a single dataset with the following structure:

- ID: a simple identifier for the entry;
- TEXT: the 4-sentence prompt to be evaluated;
- MEAN: the coherence score of the text to be predicted, based on the mean of the human judgements collected;

In the context of DisCoTEX, for both sub-tasks participants could leverage further external resources to enhance their models, with the exception of Wikipedia and mTEDx data.

4. Evaluation measures

We defined the following evaluation metrics for each sub-task:

- For sub-task 1: the evaluation metric is Accuracy (the ratio between correctly predicted samples and all processed samples) obtained by each system in the test set. We also reported Precision, Recall and F-score for the two classes;
- For sub-task 2: the evaluation metric is the harmonic mean between Pearson and Spearman correlation coefficients between the participants' scores and test set scores.

Baseline The baseline for both tasks has been computed by employing one-hot vectors representations: For sub-task 1 we extracted the one hot vector for each sentence s_i in the input prompt $P = \{s_1, s_2, ..., s_n\}$, as well as for the target sentence t. The distance between the prompt P and the target sentence t, $\mathscr{D}(P,t)$ is computed as the average distance between each sentence s_i from the prompt, and the target sentence t based on Hamming distance coefficient *Dist*:

$$\mathscr{D}(P,t) = \frac{1}{n}\left(\sum_{i=0}^{n} Dist\langle s_i, t\rangle\right).$$

To decide whether the target sentence t is coherent with the paragraph P we computed the median distance value across the whole training dataset, and we used this as a threshold: all the test samples with a distance value under the median have been considered coherent, incoherent otherwise.

For sub-task 2 we first extracted the one-hot vectors from each sentence s_i in the input prompt $P = \{s_1, s_2, ..., s_n\}$:

$$\overrightarrow{v_1} \leftarrow s_1, \overrightarrow{v_2} \leftarrow s_2, ..., \overrightarrow{v_n} \leftarrow s_n.$$

Then we computed the proximity between each consecutive vectors pair $\langle \overrightarrow{v_i}, \overrightarrow{v_{i+1}} \rangle \in V$ through Jaccard distance metric *Dist*, thereby resulting in (n-1) distance scores, grasping the degree of semantic overlap between each two neighbouring sentences. In order to compute the coherence score for the paragraph P, $\mathscr{C}(P)$, we averaged the scores featuring each pair of adjacent sentences:

$$\mathscr{C}(P) = \frac{1}{n-1}\sum_{i=1}^{n-1} Dist(\overrightarrow{v_i}, \overrightarrow{v_{i+1}}).$$

5. Participants

We received a total of 3 submissions for sub-task 1 and 2 submissions for sub-task 2 from 3 different teams. Each submission had the option to include up to three different runs. The strategies used to approach the task are all very different from each other. Teams that participated in both sub-tasks opted to use the same strategy for both challenges. None of the systems chose to utilize

Team	Members	Affiliations	sub-task 1	sub-task 2	# Runs
MPG	3	Sony Computer Science Laboratories Paris, France Enrico Fermi's Research Center (CREF), Rome, Italy, Sapienza University of Roma, Italy	✓	X	4
IUSSNets	3	Iuss Pavia, Italy	✓	✓	9
ExtremITA	4	Università degli Studi di Roma Tor Vergata, Università di Torino, Italy	✓	✓	6

Table 3
Teams participating in EVALITA 2023 DisCoTeX shared task.

additional resources apart from the official datasets. Further information regarding the task participation can be found in Table 5.

MPG

The MPG team [19] utilized the tree-based classifier LightGBM incorporating a set of explicitly engineered features aiming at comparing the prompt and target with respect to several metrics such as TF-IDF vectors, counts of upper case words, tenses, punctuation, words, and characters, as well as sentence embeddings extracted from Sentence-BERT [20]. They exclusively participated in sub-task 1 with two runs .

IUSSNets

The IUSSNets team [21] employed fine-tuning techniques on four distinct Italian language models: BERT-ita [22], Electra-ita [22], Umberto [23], and Bertino [24] separately for each sub-task. For sub-task 1, they submitted three BERT fine-tuned models: the first fine-tuned on Wikipedia (BERT 1), the second on mTEDx (BERT 2), and the third on both (BERT 3), achieving the second-place score. For sub-task 2, they submitted BERT, Bertino, and Electra fine-tuned models, once again securing the second position, primarily due to the performance of the Electra model.

ExtremITA

The ExtremITA team [25] competed using two multi-task Language Models. The first model (ExtremITA-iT5) is an encoder-decoder based on iT5-small [26], while the second model (ExtremIT-LLaMA) is a decoder based on Camoscio [27], the Italian version of LLAMA [28]. These models largely differ in number of parameters: iT5-small has approximately 110 Million parameters, while the used version of Camoscio has 7 Billion parameters. Both models underwent joint fine-tuning on all EVALITA 2023 tasks and sub-tasks, leveraging prompting techniques. For both DisCoTeX the extremIT5 model received each instance of the dataset preceeded by the task and subtask name and it produced the predicted label or score as output. Conversely, the extremITLLaMa model, which requires a structured prompt, was provided with a textual description of the task and the desired output format specification. For sub-task 1 the prompt is: *"Le due frasi precedenti, separate da '[SEP]', sono coerenti tra loro? Rispondi sì o no"*; while for sub-task 2 the prompt is: *"Quanto è coerente questa frase in una scala da 0 a 5?"*. Their team emerged as the winner across both DisCoTeX sub-tasks and datasets, thanks to the LLAMA-based model. However, the iT5-based model performed considerably worse, especially in the second sub-task where it remained below the baseline.

Team	Model	Accuracy
EXTREMITA	LLAMA	0.815
IUSSNETS	BERT	0.723
MPG	LGBM	0.595
BASELINE	HAMMING	0.525

Table 4
DisCoTeX leaderboard on sub-task 1.

6. Results

Tables 4 and 5 report the leaderboard of systems taking part in sub-task 1 and sub-task 2, respectively. Note that, for the purpose of the official ranking, for sub-task 1 we considered the accuracy of the best run, and we further computed the mean between the best result/run on Wiki and the best result/run on mTEDx data. Conversely, for sub-task 2 we first computed both Pearson and Spearman correlations, then we applied the harmonic mean between the two measures.

As it can be seen, all systems outperform the baseline in both sub-tasks. The best performance was achieved by the team extremITA with the system based on the LLAMA model.

Team	Model	r	ρ	HM
EXTREMITA	LLAMA	0.66	0.65	0.65
IUSSNETS	ELECTRA	0.65	0.62	0.63
IUSSNETS	BERT	0.64	0.60	0.62
IUSSNETS	BERTINO	0.50	0.48	0.49
BASELINE	JACCARD	0.10	0.13	0.11
EXTREMITA	T5	0.06	0.06	0.06

Table 5
DISCOTEX leaderboard on sub-task 2. Reported figures express Pearson (r) and Spearman (ρ) correlations together with their harmonic mean (HM).

7. Analysis and Discussion

To better examine the influence of genre on the automatic modeling of text coherence, we evaluated all submitted systems considering the two datasets separately. The outcomes for sub-task 1 are presented in Table 6, which displays the accuracy scores for both the positive and negative classes of the mTEDx and WIKI data, along with precision, recall, and F1 measures. From this fine-grained perspective, it appears that sub-task 1 is more effectively tackled using the dataset extracted from Wikipedia, with an average accuracy of 0.71, compared to the subset extracted from mTEDx talks, which achieved an average accuracy of 0.63. This discrepancy in performance could be attributed to the structural differences in the texts. Language models indeed are more frequently exposed to the encyclopedic language found in Wikipedia, whereas the lecture-style transcriptions of spoken language in mTEDx talks pose a greater challenge. Furthermore, determining whether the target sentence directly follows the prompt may be more difficult in the transcription of a mTEDx talk. Indeed, unlike traditional lectures, mTEDx talks are a type of public speech that is designed to engage a broad audience by exploring a single idea with unwavering focus. Focusing on the single submissions, the extremITA team outperforms the other participants on both datasets with the run corresponding to the LLaMA based model. Such improvement over competitors may be due to the large amount of model parameters (7B) together with the multi-task setting. The results obtained by the IUSSNets team are comparable to those obtained by the first run of extremITA: although both models are based on Transformers architecture, the results obtained by the first and second runs on mTEDx and Wiki by the IUSSNets team seem to be more stable in terms of Precision and Recall. Regarding the MPG team's results, the higher difference from other competing models may indicate the lack of discourse-level features, these could be beneficial to fully grasp the properties of the word sequence.

The detailed results for the sub-task 2 are reported in Table 7. In contrast to the first sub-task, in the second one

Team	Model (run)	Ted						
		0			1			Accuracy
		P	R	F1	P	R	F1	
IUSSNETS	BERT(1)	0,71	0,70	0,70	0,70	0,71	0,71	**0,70**
	BERT(2)	—	—	—	—	—	—	—
	BERT(3)	0,50	0,28	0,36	0,50	0,71	0,59	0,50
MPG	LGBM(1)	0,54	0,78	0,64	0,60	0,32	0,42	0,55
	LGBM(2)	0,55	0,67	0,60	0,57	0,45	0,50	**0,56**
EXTREMITA	T5	0,77	0,57	0,66	0,66	0,83	0,74	0,70
	LLAMA	0,78	0,80	0,79	0,79	0,78	0,79	**0,79**
BASELINE	HAMMING	0,51	0,43	0,47	0,51	0,59	0,54	0,51

Team	Run ID	Wiki						
		0			1			Accuracy
		P	R	F1	P	R	F1	
IUSSNETS	BERT(1)	—	—	—	—	—	—	—
	BERT(2)	0,75	0,71	0,73	0,72	0,76	0,74	**0,74**
	BERT(3)	—	—	—	—	—	—	—
MPG	LGBM(1)	0,61	0,69	0,65	0,64	0,56	0,60	0,62
	LGBM(2)	0,63	0,64	0,63	0,63	0,62	0,62	**0,63**
EXTREMITA	T5	0,85	0,49	0,62	0,64	0,91	0,75	0,70
	LLAMA	0,87	0,81	0,84	0,82	0,88	0,85	**0,84**
BASELINE	HAMMING	0,54	0,50	0,52	0,54	0,58	0,56	0,54

Table 6
Detailed results on sub-task 1 considering mTEDx (top) and Wikipedia (bottom) datasets.

the best results are generally obtained on the subset of mTEDx talks. Indeed, the system submitted by IUSSNets seems to suffer particularly from the Wiki subset: the third run of IUSSNets has a difference of 0.24 and 0.27 in terms of Pearson and Spearman correlations respectively, while the second run of extremITA only 0.12 and 0.11 respectively. The unexpected best performance of systems on mTEDx talks warrants further investigation, especially considering the higher variance in coherence scores obtained for prompts extracted from TED talks (std: 0.92) than those extracted from Wiki data (std: 0.63).

Team	Model (run)	Ted		Wiki		Pert.		Non-Pert.	
		r	ρ	r	ρ	r	ρ	r	ρ
EXTREMITA	T5	0,12	0,13	0,03	0,01	0,06	0,07	0,08	0,04
	LLAMA	0,65	0,66	**0,53**	**0,55**	**0,65**	0,66	**0,66**	**0,6**
IUSSNETS	BERTINO	0,46	0,48	0,31	0,3	0,53	0,55	0,48	0,35
	BERT	0,66	0,68	0,44	0,44	0,63	0,64	0,66	0,51
	ELECTRA	**0,69**	**0,71**	0,45	0,44	**0,65**	**0,69**	**0,66**	0,49
BASELINE	JACCARD	0,12	0,18	0,13	0,16	0,08	0,14	0,11	0,09

Table 7
Detailed results on sub-task 2, in terms of Pearson (r) and Spearman (ρ) correlations for both subsections. We additionally reported correlation scores for perturbed (Pert.) data and unperturbed data (Non-Pert.).

8. Conclusion

We presented the results of the DISCOTEX task, held within EVALITA 2023 [14]. The task was divided in two sub-tasks: the first one challenged participants to propose systems able to discriminate between coherent and incoherent textual passages, where the latter have been artificially created to gradually undermine local coher-

ence within text. The second one intended to model the human perception of text coherence by predicting the average score attributed to human raters to a text. A novel dataset was developed for this task comprising texts from two different domains, representative of a written and spoken language variety in order to investigate the role of modality on the automatic modeling of coherence. Three teams participated in the task and submitted a total of 19 runs. Notably, the ExtremITA team secured the first position in both sub-tasks with their system based on the largest decoder model proposed. However, it is worth highlighting that smaller models with fewer parameters also demonstrated comparable performance, indicating their effectiveness in capturing discourse-related information. Quite surprisingly, the results of sub-task 2 revealed that systems were more proficient in predicting coherence scores for TEDx talks compared to Wikipedia texts, which calls for further investigation by also expanding the current dataset of human evaluated texts. Future plans involve extending the DisCoTeX task to a multilingual perspective, enabling coherence modeling exploration across different languages using reproducible data collection processes in languages with available Wiki and TED resources.

Acknowledgements

The authors gratefully acknowledge the support of the PNRR MUR project PE0000013-FAIR.

References

[1] T. A. Van Dijk, W. Kintsch, Strategies of discourse comprehension, Academic Press, New York, 1983.

[2] B. J. Grosz, A. K. Joshi, S. Weinstein, Centering: A framework for modeling the local coherence of discourse, Computational Linguistics 21 (1995) 203–225. URL: https://aclanthology.org/J95-2003.

[3] R. Barzilay, M. Lapata, Modeling Local Coherence: An Entity-Based Approach, Computational Linguistics 34 (2008) 1–34. URL: https://doi.org/10.1162/coli.2008.34.1.1. doi:10.1162/coli.2008.34.1.1. arXiv:https://direct.mit.edu/coli/article-pdf/34/1/1/1798481/coli.2008.34.1.1.pdf.

[4] M. Elsner, E. Charniak, Disentangling chat with local coherence models, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Portland, Oregon, USA, 2011, pp. 1179–1189. URL: https://aclanthology.org/P11-1118.

[5] D. Tien Nguyen, S. Joty, A neural local coherence model, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 1320–1330. URL: https://aclanthology.org/P17-1121. doi:10.18653/v1/P17-1121.

[6] J. Li, D. Jurafsky, Neural net models of open-domain discourse coherence, ArXiv abs/1606.01545 (2017).

[7] A. Lai, J. R. Tetreault, Discourse coherence in the wild: A dataset, evaluation and methods, CoRR abs/1805.04993 (2018). URL: http://arxiv.org/abs/1805.04993. arXiv:1805.04993.

[8] M. Mesgar, M. Strube, A neural local coherence model for text quality assessment, in: Proceedings of the 2018 conference on empirical methods in natural language processing, 2018, pp. 4328–4339.

[9] B. Elvevåg, P. W. Foltz, D. R. Weinberger, T. E. Goldberg, Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia, Schizophrenia research 93 (2007) 304–316.

[10] D. Iter, J. Yoon, D. Jurafsky, Automatic detection of incoherent speech for diagnosing schizophrenia, in: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, 2018, pp. 136–146.

[11] A. Shen, M. Mistica, B. Salehi, H. Li, T. Baldwin, J. Qi, Evaluating Document Coherence Modeling, Transactions of the Association for Computational Linguistics 9 (2021) 621–640. URL: https://doi.org/10.1162/tacl_a_00388. doi:10.1162/tacl_a_00388.

[12] M. Chen, Z. Chu, K. Gimpel, Evaluation benchmarks and learning criteria for discourse-aware sentence representations, arXiv preprint arXiv:1909.00142 (2019).

[13] Y. Farag, J. Valvoda, H. Yannakoudakis, T. Briscoe, Analyzing neural discourse coherence models, in: Proceedings of the First Workshop on Computational Approaches to Discourse, Association for Computational Linguistics, Online, 2020, pp. 102–112. URL: https://aclanthology.org/2020.codi-1.11. doi:10.18653/v1/2020.codi-1.11.

[14] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, EVALITA 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[15] D. Brunato, C. Chesi, F. Dell'Orletta, S. Montemagni, G. Venturi, R. Zamparelli, AcCompl-it @ EVALITA2020: Overview of the Acceptability & Complexity Evaluation Task for Italian, EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020 (2020).

[16] L. Gregori, M. Montefinese, D. P. Radicioni, A. A.

Ravelli, R. Varvara, Concretext @ EVALITA2020: The Concreteness in Context Task., EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020 (2020).
[17] G. Caliendo, The popularisation of science in web-based genres, The language of popularisation: Theoretical and descriptive models 3 (2012) 101–132.
[18] D. Biber, S. Conrad, R. Reppen, Corpus linguistics: investigating language structure and use., Cambridge University Press, Cambridge, 1998.
[19] M. Galletti, P. Gravino, G. Prevedello, MPG at DisCoTex: Predicting text coherence by tree-based modelling of linguistic features, in: M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi (Eds.), Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, September 7th-8th 2023, Parma, 2023.
[20] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. URL: https://aclanthology.org/D19-1410. doi:10.18653/v1/D19-1410.
[21] E. Zanoli, M. Barbini, C. Chesi, IussNets at DisCoTex: A fine-tuned approach to coherence, in: M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi (Eds.), Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, September 7th-8th 2023, Parma, 2023.
[22] S. Schweter, Italian BERT and ELECTRA models, 2020. URL: https://doi.org/10.5281/zenodo.4263142. doi:10.5281/zenodo.4263142.
[23] L. Parisi, S. Francia, P. Magnani, UmBERTo: an Italian language model trained with whole word masking, https://github.com/musixmatchresearch/umberto, 2020.
[24] M. Muffo, E. Bertino, BERTino: an Italian DistilBERT model, Computational Linguistics CLiC-it 2020 (2020) 317.
[25] C. D. Hromei, D. Croce, V. Basile, R. Basili, ExtremITA at EVALITA2023: Multi-task sustainable scaling to large language models at its extreme, in: M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi (Eds.), Proceedings of Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, September 7th-8th 2023, Parma, 2023.
[26] G. Sarti, M. Nissim, IT5: Large-scale text-to-text pretraining for Italian language understanding and generation, ArXiv preprint 2203.03759 (2022). URL: https://arxiv.org/abs/2203.03759.
[27] A. Santilli, Camoscio: An Italian instruction-tuned LLaMa, https://github.com/teelinsan/camoscio, 2023.
[28] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, LLaMa: Open and efficient foundation language models, 2023. arXiv:2302.13971.

IUSSNets at DisCoTeX: A fine-tuned approach to coherence

Emma Zanoli[1], Matilde Barbini[1] and Cristiano Chesi[1]

[1]University School for Advanced Studies IUSS Pavia - NeTS Lab

Abstract

We present our submission to the DisCoTex shared task of the EVALITA 2023 evaluation campaign, which focuses on modeling discourse coherence for Italian texts. We highlight the importance of coherence modeling in natural language processing tasks and briefly discuss related work, including earlier linguistic theories and recent neural models. To tackle the task, we leverage pre-trained Transformer models and fine-tune them on the provided datasets. Our approach incorporates monolingual models due to limited computing resources, but shows potential for multilingual and multitask learning. Our systems ranks second overall, showing that Transformer models can be fruitfully leveraged for coherence assessment, but more work is needed to fully exploit their capabilities. The coherence assessment literature focuses primarily on English; this shared task and our work contribute to broadening the scope of current research.

Keywords

coherence, Transformers, NLP, computational linguistics

1. Introduction

Written texts are often a sequence of semantically coherent segments, designed to create a smooth transition between various subtopics [1]. Modeling coherence can be done by building text analysis models that can distinguish a coherent text from incoherent ones, or that can output a coherence score [2]. It has been a key problem in discourse analysis, with applications in many downstream NLP tasks (e.g. text generation, summarization, machine translation, dialogue generation, etc.).

Coherence modeling is at the heart of the DisCoTex shared task [3] of the EVALITA 2023 evaluation campaign [4]. This report relates the motivation and implementation of the IUSSnets team's submission.

2. Related work

Early computational models for text coherence assessment were mainly based on one of two linguistic theories: a) centering theory [5] and b) rhetorical structure theory [6]. In line with the first, [7] and [8] use the distribution of entity transitions over sentences to predict text coherence. In line with the second, [9] and [10] produce discourse relations over sentences with a discourse parser, showing that the relations are indicative of text coherence.

More recently, neural models have gained prominence in the task of coherence assessment. Popular examples are [11], [12], [13], and the recent state-of-the-art [14].

Our implementation choices are informed by [15], who are among the first to use Transformer models for coherence assessment.

It is interesting to note that the literature on coherence finds significant overlap with the literature on readability. The two are often likened and used as general measures of textual quality [9]. Sometimes, coherence is used as an additional feature in readability assessment [12].

By and large, the literature on automatic assessment of discourse coherence focuses on the English language. One notable exception is [16] for Danish.

3. Task

DisCoTEX is the first shared task focused on modelling discourse coherence for Italian real-word texts. The organizers proposed two sub-tasks:

- **Sub-task 1 - Last sentence classification**: a binary classification task. Given a short paragraph (the prompt), and an individual sentence (the target), the goal is to classify whether the target follows or not, i.e. whether joining it to the prompt gives out a coherent or incoherent text.
- **Sub-task 2 - Human score prediction**: a regression task. The goal is to predict the average coherence score assigned by human raters to short paragraphs. Judgments are expressed on a 5-point Likert Scale.

4. DisCoTex Data

The dataset for the DisCoTex task contains texts extracted from two sources: the Italian Wikipedia and the section of Italian speech transcripts included in the Multilingual TEDx corpus (mTEDx). For both subtasks, coherence is

Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023)

emma.zanoli@iusspavia.it (E. Zanoli);
matilde.barbini@iusspavia.it (M. Barbini);
cristiano.chesi@iusspavia.it (C. Chesi)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper50.pdf

analyzed within text passages of four consecutive sentences. For task 1, these were split into 8000 prompt-target pairs for each domain: the prompt is always made of the first three consecutive sentence, whereas the target can either be the actual last sentence of the passage (for the positive class) or a different one (for the negative class). This dataset is automatically generated. For task 2 there were 1064 text passages, equally balanced across the two original source datasets, of which 50% were left unaltered and 50% were artificially modified to undermine coherence. This dataset was not automatically generated: each passage was annotated by at least 10 human evaluators who were native speakers of Italian.

5. Description of the system

5.1. General intuition

For this challenge we leveraged pre-trained Transformer models and fine-tuned them on the provided data.

Transformer models [17] have been applied with tremendous success to the field of NLP. They have been shown to capture semantic relationships to a reasonable extent. As reported in Section 2, they have already successfully been applied to the task of discourse coherence modeling.

Since the DisCoTex task is tailored specifically to the Italian language, we decided to leverage monolingual Transformer models that had been pre-trained exclusively on Italian data. Given that coherence assessment datasets are available for English, we initially intended to experiment with multi-lingual transfer learning, using multilingual pre-trained Transformer models and fine-tuning them simultaneously on English and Italian data. Unfortunately, our limited computing resources did not allow us to get this far within the time frame of the shared task. Preliminary results indicate that this would have been a promising approach.

5.2. Pre-trained models

We experimented with 4 monolingual pre-trained models, freely available on the HuggingFace hub [18] at the time of writing:

- `bert-ita`[1]: an Italian version of BERT [19];
- `electra-ita`[2]: an Italian version of ELECTRA [19];
- `umberto`[3]: an Italian version of RoBERTa [20];
- `bertino`[4]: an Italian version of DistilBERT [21].

[1] https://huggingface.co/dbmdz/bert-base-italian-xxl-cased
[2] https://huggingface.co/dbmdz/electra-base-italian-xxl-cased-discriminator
[3] https://huggingface.co/Musixmatch/umberto-commoncrawl-cased-v1
[4] https://huggingface.co/indigo-ai/BERTino

In the following we provide an overview of the main intuition for each model.

BERT by Google [22] introduced "masked language modeling" (MLM): some of the input tokens were masked, and the pre-training objective was to predict the original vocabulary id of the masked word based only on its context. MLM enabled the representation to fuse the left and the right context, leading to a bidirectional Transformer. In addition to MLM, they also used a "next sentence prediction" task that jointly pre-trained text-pair representations. After pre-training, BERT could be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, without substantial task-specific architecture modifications.

DistilBERT by HuggingFace [23] leveraged knowledge distillation during the pre-training phase, thus reducing the size of a BERT model by 40%, while retaining 97% of its language understanding capabilities and being 60% faster. To leverage the inductive biases learned by larger models during pre-training, they introduced a triple loss combining language modeling, distillation and cosine-distance losses.

RoBERTa by Facebook AI [24] applied various pre-training enhancements to the original BERT model: longer training on longer sequences, bigger batches over more data, no next sentence prediction objective, and dynamically changing the masking pattern applied to the training data. These modifications advanced the state of the art on different downstream tasks.

ELECTRA by Stanford and Google [25] introduced a new pre-training task called "replaced token detection": instead of masking the input, they corrupted it by replacing some tokens with plausible alternatives sampled from a small generator network. Then, instead of predicting the original identities of the corrupted tokens, they trained a discriminative model that predicts whether each token in the corrupted input was replaced by a generator sample or not. The model showed competitive performance compared to other models, while requiring fewer resources for training.

As previously stated, we experimented with monolingual Italian versions of these models, i.e. models that were trained using the same approaches as the ones described above, but solely on Italian data. These models were used to encode the input and return a vector representation from the last layer output (i.e. the `[CLS]` token, which was taken to signify a vector representation of the sentence).

5.3. Fine-tuning

The pre-trained models were fine-tuned on the available data for 10 epochs, using the following hyper-parameters: 0.1 dropout rate, 0.01 weight decay, 1e-6 learning rate, a batch size of 1 and no gradient clipping. We used the cur-

Model	Training data	Language	Dataset size
`bert-ita`	Wikipedia, OPUS [27], OSCAR [28]	Italian	81 GB
`electra-ita`	Wikipedia, OPUS [27], OSCAR [28]	Italian	81 GB
`umberto`	OSCAR [28] - deduplicated	Italian	70 GB
`bertino`	PAISÀ [29], ItWaC [30]	Italian	12 GB

Table 1
The pre-trained models we use and the datasets they were trained on.

Team	Run ID	Ted							Wiki						
		0			1			Accuracy	0			1			Accuracy
		P	R	F1	P	R	F1		P	R	F1	P	R	F1	
IUSSnets	run1	0.71	0.70	0.70	0.70	0.71	0.71	**0.70**	—	—	—	—	—	—	—
	run2	—	—	—	—	—	—	—	0.75	0.71	0.73	0.72	0.76	0.74	**0.74**
	run3	0.50	0.28	0.36	0.50	0.71	0.59	0.50	—	—	—	—	—	—	—
baseline	Hamming	0.51	0.43	0.47	0.51	0.59	0.54	0.51	0.54	0.50	0.52	0.54	0.58	0.56	0.54

Table 2
Full official results, sub-task 1. Using `bert-ita` for all 3 runs. run1: trained on the ted dataset; run2: trained on the wiki dataset; run3: trained on both datasets.

rently available PyTorch implementation of the Adam optimizer [26], `torch.optim.Adam`. During fine-tuning, the embedding layers of the pre-trained models were frozen.

5.4. Data

During fine-tuning, we only relied on the provided datasets. However, we used Transformer models which had been pre-trained on a variety of data sources (see Table 1).

For sub-task 2 we attempted some data augmentation techniques. Since we had a dataset were each sentence had a mean score based on at least 10 judgments, we leveraged the standard deviation to generate a distribution of 10 scores that would have the provided mean and standard deviation. We thus ended up with 10 scores for each sentence, instead of an average score. However, upon training our models on this augmented dataset, we did not notice any significant improvements and, because this approach was more resource-intensive, we eventually dropped it.

Please note that we only made use of 80% of the provided datasets during fine-tuning; the remaining 20% was used as a validation split (more details below).

6. Results

For the purposes of the official rankings, our results are: **0.72** on sub-task 1, **0.63** on sub-task 2.

For sub-task 1, the organizers considered the accuracy of the best run and computed the mean between the best results on the two datasets (Ted and Wiki). For sub-task 2, they first computed both Pearson and Spearman correlations, then they applied the harmonic mean between the two measures. Participants were allowed to submit up to 3 runs per sub-task; the full official results of our submission can be seen in Table 2 and Table 3.

For several months during training and experimentation, we were not made privy to the exact way the performance of our models would be calculated. The task instructions specified that task 1 would be evaluated according to accuracy and a second metric (which was never disclosed), whereas task 2 would be evaluated with a metric based on a standard correlation coefficient ("Pearson and/or Spearman" - it ended up being a harmonic mean of the two). During our experimentation, we decided to evaluate on accuracy for task 1, and on Spearman correlation for task 2. The two sections below report the respective results.

6.1. Sub-task 1 - evaluation results

In the absence of a test or validation set, we sampled 20% of the original training sets for preliminary evaluation. This resulted in 1600 randomly sampled data points for each dataset. On these sub-sets, we calculated the binary accuracy as implemented in the `torchmetrics` Python library[5]. We report results in Table 4.

6.2. Sub-task 2 - evaluation results

In the absence of a test or validation set, we sampled 20% of the original training set for preliminary evaluation. This resulted in 172 randomly sampled data points. On this sub-set, we computed the Spearman correlation coefficient as implemented in the `scipy` Python library[6]. We report results in Table 5.

[5] https://torchmetrics.readthedocs.io/en/stable/classification/accuracy.html#binaryaccuracy
[6] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html

Team	Run ID	Pearson Corr.	Spearman Corr.	Harm. Mean
IUSSnets	run1	0.50	0.48	0.49
	run2	0.64	0.60	0.62
	run3	0.65	0.62	**0.63**
baseline	Jaccard	0.10	0.13	0.11

Table 3
Full official results, sub-task 2. run1: `bertino`; run2: `bert-ita`; run3: `electra-ita`.

Model	Dataset	Accuracy
`bert-ita`	wiki	**0.749**
`electra-ita`	wiki	0.716
`umberto`	wiki	0.595
`bertino`	wiki	0.637
`bert-ita`	all	**0.723**
`electra-ita`	all	0.583
`bert-ita`	ted	**0.704**
`electra-ita`	ted	0.617

Table 4
Evaluation results, sub-task 1.

Model	Spearman Corr.
`bert-ita`	0.574
`electra-ita`	**0.637**
`umberto`	0.464
`bertino`	0.562

Table 5
Evaluation results, sub-task 2.

7. Discussion

The DisCoTex shared task provided us with an excellent opportunity to reflect on the notion of disourse coherence and on the ways it may be assessed, whether automatically or not.

As a preamble, let us note that datasets for coherence assessment that are automatically created by shuffling existing texts have been criticized, among others, by [31] and [32], and the models trained on them have been shown to perform weakly on downstream tasks [2]. Nonetheless, such datasets have remained common benchmarks.

Discourse coherence is a complicated concept that is related to almost every aspect of discourse communication. In the linguistics literature, there is no all-embracing rule governing coherence analysis: different scholars have presented their insight into different aspects of discourse coherence [33]. When we read a text or listen to speech, we are inclined to infuse it with coherence by making our own inferences based on our understanding and perception. Coherence is therefore achieved not by using superficial markers such as linguistic or grammatical devices, but through psychological, cognitive, or pragmatic means. The comprehension of discourse and an appreciation for its coherence are driven by active inference, background knowledge, and a degree of imagination.

It comes as no surprise, then, that the many facets of this uniquely human experience are hard to model computationally. In order to get a sense for this, we looked into the dataset collected for sub-task 2. Overall, the majority of the training dataset contained texts rated 3.0 or higher; in other words, the texts were perceived as mostly coherent. It would have been interesting to compare how the annotators rated original vs. artificially modified text passages. Although we did not have this information in the dataset, when comparing the datasets for sub-task 1 and 2, we found 19 passages in the dataset for sub-task 2 in the positive class of the ted dataset for sub-task 1: this means that these passages had not been modified from their original sources and were thus expected to be coherent. Of these 19 passages:

- none were unanimously rated as coherent, i.e. a mean score of 5 (0%);
- 4 received a mean score of 4 or above (21%);
- 10 received a rating between 3 and 4 (53%);
- 4 received a rating between 2 and 3 (21%);
- 1 even received a rating below 2 (5%).

If we were to revert these scores back to a binary classification (with a halfway cutoff at 2.5), 5 of these passages would be considered incoherent. However, for the purposes of sub-task 1, they would have been considered coherent. This simplistic example is in no way an exhaustive exploration of the nature of the tasks or the provided datasets, but it serves the purpose of reflecting on the difficulty of modeling these phenomena from a more explicit (linguistic or cognitive) perspective.

Deep learning models generally, and Transformers specifically, have been shown to capture useful semantic information in texts. Previous work has investigated Transformers for their semantic [34] and even pragmatic [35] properties. For these reasons, we hypothesized that Transformer models would be a good fit for the task of coherence assessment. Indeed, even in our simple setup, we can see promising results. Further experimentation and greater computational power could lead to significant performance improvements. Multilingual and multi-task learning might prove particularly effective in boosting performance on Italian texts by leveraging datasets that exist for the English language or for other related tasks.

Moving forward, further exploration of linguistic theories and neural models can enhance discourse coherence assessment and facilitate more sophisticated language processing applications. Focusing on more controlled textual continuations (e.g. different logical conclusions from specific premises) would shed some light on the relevance of specific factors in coherence modeling. This would also allow us to better understand the strengths and weaknesses of a transformers-based approach.

Acknowledgments

We are thankful to the bright research community at the NeTS lab of IUSS Pavia, who encouraged and supported these experiments. This research has been partially founded by the PON Governance 2014-2020: Next Generation UPP Project - CUP D19J22000240006.

References

[1] D. Aumiller, S. Almasian, S. Lackner, M. Gertz, Structural text segmentation of legal documents, in: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law, ICAIL '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 2–11. URL: https://doi.org/10.1145/3462757.3466085. doi:10.1145/3462757.3466085.

[2] T. Mohiuddin, P. Jwalapuram, X. Lin, S. Joty, Rethinking coherence modeling: Synthetic vs. downstream tasks, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 3528–3539. URL: https://aclanthology.org/2021.eacl-main.308. doi:10.18653/v1/2021.eacl-main.308.

[3] D. Brunato, D. Colla, F. Dell'Orletta, I. Dini, D. P. Radicioni, A. A. Ravelli, DisCoTex at EVALITA 2023: Overview of the Assessing DIScourse COherence in Italian TEXts task, in: M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi (Eds.), Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[4] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[5] B. J. Grosz, A. K. Joshi, S. Weinstein, Centering: A framework for modeling the local coherence of discourse, Computational Linguistics 21 (1995) 203–225. URL: https://aclanthology.org/J95-2003.

[6] W. C. Mann, S. A. Thompson, Rhetorical structure theory: Toward a functional theory of text organization, Text-interdisciplinary Journal for the Study of Discourse 8 (1988) 243–281.

[7] R. Barzilay, M. Lapata, Modeling local coherence: An entity-based approach, Computational Linguistics 34 (2008) 1–34.

[8] C. Guinaudeau, M. Strube, Graph-based local coherence modeling, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Sofia, Bulgaria, 2013, pp. 93–103. URL: https://aclanthology.org/P13-1010.

[9] E. Pitler, A. Nenkova, Revisiting readability: A unified framework for predicting text quality, in: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Honolulu, Hawaii, 2008, pp. 186–195. URL: https://aclanthology.org/D08-1020.

[10] Z. Lin, H. T. Ng, M.-Y. Kan, Automatically evaluating text coherence using discourse relations, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Portland, Oregon, USA, 2011, pp. 997–1006. URL: https://aclanthology.org/P11-1100.

[11] D. Tien Nguyen, S. Joty, A neural local coherence model, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 1320–1330. URL: https://aclanthology.org/P17-1121. doi:10.18653/v1/P17-1121.

[12] M. Mesgar, M. Strube, A neural local coherence model for text quality assessment, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 4328–4339. URL: https://aclanthology.org/D18-1464. doi:10.18653/v1/D18-1464.

[13] H. C. Moon, T. Mohiuddin, S. Joty, C. Xu, A unified neural coherence model, 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 2262–2272. URL: https://aclanthology.org/D19-1231. doi:10.18653/v1/D19-1231.

[14] M. Mesgar, L. F. R. Ribeiro, I. Gurevych, A neu-

ral graph-based local coherence model, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 2316–2321. URL: https://aclanthology.org/2021.findings-emnlp.199. doi:10.18653/v1/2021.findings-emnlp.199.

[15] T. Abhishek, D. Rawat, M. Gupta, V. Varma, Transformer models for text coherence assessment, 2022. arXiv:2109.02176.

[16] L. Flansmose Mikkelsen, O. Kinch, A. Jess Pedersen, O. Lacroix, DDisCo: A discourse coherence dataset for Danish, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 2440–2445. URL: https://aclanthology.org/2022.lrec-1.260.

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

[18] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Brew, Huggingface's transformers: State-of-the-art natural language processing, ArXiv abs/1910.03771 (2019).

[19] S. Schweter, Italian bert and electra models, 2020. URL: https://doi.org/10.5281/zenodo.4263142. doi:10.5281/zenodo.4263142.

[20] L. Parisi, S. Francia, P. Magnani, Umberto: an italian language model trained with whole word masking, https://github.com/musixmatchresearch/umberto, 2020.

[21] M. Muffo, E. Bertino, Bertino: an italian distilbert model, https://github.com/indigo-ai/BERTino, 2020.

[22] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[23] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2019. arXiv:1910.01108.

[24] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. arXiv:1907.11692.

[25] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, ELECTRA: Pre-training text encoders as discriminators rather than generators, in: ICLR, 2020. URL: https://openreview.net/pdf?id=r1xMH1BtvB.

[26] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2017. arXiv:1412.6980.

[27] J. Tiedemann, Parallel data, tools and interfaces in opus, in: N. C. C. Chair), K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis (Eds.), Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey, 2012.

[28] P. J. Ortiz Suárez, B. Sagot, L. Romary, Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures, Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, Leibniz-Institut für Deutsche Sprache, Mannheim, 2019, pp. 9 – 16. URL: http://nbn-resolving.de/urn:nbn:de:bsz:mh39-90215. doi:10.14618/ids-pub-9021.

[29] V. Lyding, E. Stemle, C. Borghetti, M. Brunello, S. Castagnoli, F. Dell'Orletta, H. Dittmann, A. Lenci, V. Pirrelli, The PAISÀ corpus of Italian web texts, in: Proceedings of the 9th Web as Corpus Workshop (WaC-9), Association for Computational Linguistics, Gothenburg, Sweden, 2014, pp. 36–43. URL: https://aclanthology.org/W14-0406. doi:10.3115/v1/W14-0406.

[30] M. Baroni, S. Bernardini, A. Ferraresi, E. Zanchetta, The wacky wide web: a collection of very large linguistically processed web-crawled corpora, Language resources and evaluation 43 (2009) 209–226.

[31] P. Laban, L. Dai, L. Bandarkar, M. A. Hearst, Can transformer models measure coherence in text: Rethinking the shuffle test, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics, Online, 2021, pp. 1058–1064. URL: https://aclanthology.org/2021.acl-short.134. doi:10.18653/v1/2021.acl-short.134.

[32] A. Beyer, S. Loáiciga, D. Schlangen, Is incoherence surprising? targeted evaluation of coherence prediction from language models, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 4164–4173. URL: https://aclanthology.org/2021.naacl-main.328. doi:10.18653/v1/2021.naacl-main.328.

[33] Y. Wang, M. Guo, A short analysis of discourse coherence, Journal of Language Teaching and Research 5 (2014) 460.

[34] E. Reif, A. Yuan, M. Wattenberg, F. B. Viegas, A. Coenen, A. Pearce, B. Kim, Visualizing and measuring the geometry of bert, Advances in Neural Information Processing Systems 32 (2019).

[35] L. Pandia, Y. Cong, A. Ettinger, Pragmatic competence of pre-trained language models through the lens of discourse connectives, in: Proceedings of the 25th Conference on Computational Natural Language Learning, Association for Computational Linguistics, Online, 2021, pp. 367–379. URL: https://aclanthology.org/2021.conll-1.29. doi:10.18653/v1/2021.conll-1.29.

MPG at DisCoTeX: Predicting Text Coherence by Tree-based Modelling of Linguistic Features

Martina Galletti[1,2,3], Pietro Gravino[1,2] and Giulio Prevedello[1,2]

[1]Sony Computer Science Laboratories Paris, 6, Rue Amyot, 75005, Paris, France
[2]Enrico Fermi's Research Center (CREF), via Panisperna 89A, 00184, Rome, Italy
[3]Department of Computer, Control and Management Engineering (DIAG) "Antonio Ruberti", Sapienza University of Rome, via Ariosto 25, Rome, 00185, Italy

Abstract
Automatic text coherence modelling plays a crucial role in natural language processing tasks, such as machine translation, summarisation, and question answering. Moreover, text coherence is fundamental to reading comprehension and readers' engagement, essential to a number of application domains. In this report, we report progress for the Assessing Discourse Coherence in Italian Texts task from EVALITA-23, whose goal is to address automatic coherence detection. The task was challenged by extracting linguistic features used to train a machine learning classifier, leading to minor improvement over the baseline. The feature importance analysis revealed semantic features' relevance, providing indications for future feature engineering and modelling efforts.

Keywords
Natural language processing, text coherence, Italian language, machine learning

1. Introduction

Coherence is an essential quality to facilitate comprehension and readers' engagement. Text coherence measures the quality of organization in the structure of a text and the extent to which a reader can follow the relationships between sentences and paragraphs. Several text coherence models exist in the literature which aim to distinguish between coherent and incoherent documents. This distinction significantly affects some downstream tasks, such as document summarisation, automatic essay assessment, and the creation of natural language [1, 2, 3]. Traditional approaches often focus on linguistic features such as discourse markers [4, 5], textual connectives [6], entity-grid based approaches [7, 8] which takes inspiration from Centering Theory [9] to capture coherence [7, 10]. Recent approaches rely on neural architectures and use Convolutional Neural Networks (CNNs) over an entity-based representation of text [11, 12], Sequence to Sequence Models [3, 13] and Multi-Task Learning [14]. Nevertheless, language models still face challenges in capturing and predicting global coherence across longer texts, and targeted evaluation paradigms are still being implemented [15, 16, 17, 18, 19].

In addition to being fundamental in a number of NLP tasks, the automatic assessment of coherence in a text could have significant applications in hybrid speech and language therapy. Coherence evaluation can assess various text sources, such as educational materials, therapy resources, or online content, commonly used in therapy sessions. By utilizing automatic metrics, therapists can have measures to judge the complexity and accessibility of these texts faster, selecting appropriate materials that align with their patients' reading abilities. This ensures therapy resources are tailored to individual needs, promoting engagement, comprehension, and progress.

In the context of news media, coherence measurement is important to evaluate news items' readability but is also involved in the assessment of the alignment between an article's content and headline. Sensationalized or exaggerated headlines are sometimes used to attract attention and gain clicks to online news, a practice termed clickbaiting that undermines trust in news media. The automatic detection of text coherence can empower readers to focus on accurate journalistic products, thus contributing to the transparency and trust in news media and encouraging responsible reporting of information.

In this work, text coherence is addressed by evaluating whether a target sentence follows a small prompt paragraph (made of a few sentences) in a coherent (or not) way. This problem is a subtask from the challenge "Assessing Discourse Coherence in Italian Texts" (DisCoTeX) [20] presented at EVALITA 2023 [21]. In this report, the feature extraction procedure and the modelling strategy are described in Section 2, while the results are illustrated in Section 3 and discussed in Section 4.

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, September 07–08, 2023, Parma, Italy
martina.galletti@sony.com (M. Galletti); pietro.gravino@sony.com (P. Gravino); giulio.prevedello@sony.com (G. Prevedello)
0000-0002-0937-8830 (P. Gravino); 0000-0002-9857-2351 (G. Prevedello)
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3473/paper51.pdf

2. Description of the system

The data provided were cleaned and pre-processed in a series of systematic steps. After removing irrelevant characters, the dataset provided by the task organisers was normalised in a standardised format, i.e. lemmas, to break the text into meaningful units. Then, words with less than three characters were removed, keeping stop words. This was done because short words often carry less semantic meaning than longer words, and thus they could increase the vocabulary size of a model without contributing significantly to its training. By removing these words, some potential noise in the training could be removed. On the other hand, stop words, especially conjunctions, were kept for their role in connecting sentences across a text and preserving syntactic and grammatical relationships during the training phase, even if they do not carry a semantic meaning *per se*. After lemmatization, we added more info to the provided data by computing the length of words and sentences. This was done because the length of the words and/or sentences can provide insights into the complexity and readability of the text, which can, in turn, impact its intrinsic coherence. Longer words can indicate, in fact, a more complex and domain-specific vocabulary, while longer sentences could indicate the presence of multiple sub-clauses, which could impact the overall coherence. Moreover, an abrupt variation in the length of words and sentences could indicate an unbalance in the structure of the text, thus endangering its linguistic coherence. For similar reasons, the statistics of the uses of the different tenses in sentences were also computed since the usage of appropriate tenses ensures temporal consistency and logical progression of information. Afterwards, we extracted lexical features such as word frequency, for which we used document term matrix, Term Frequency - Inverse Document Frequency (TF-IDF), and sentence embeddings, i.e. Sentence-BERT (SBERT) [22].

To compress high-dimensional vectors, from the TF-IDF analysis and the sentence embedding, Uniform Manifold Approximation and Projection (UMAP) was used to reduce their dimension down to 30 components [23]. UMAP was chosen over the principal component analysis method as it tends to preserve local distances better. Meanwhile, compared to the t-distributed stochastic neighbour embedding method, it is faster and better preserves the global data structure.

Finally, prompt and target were compared by means of the statistics mentioned above, resulting in the following list of features for each data point:

- *tfidf_raw_wjac*: weighted Jaccard distance between prompt's and target's TF-IDF vectors;
- *tfidf_raw_cos*: cosine distance between prompt's and target's TF-IDF vectors;
- *tfidf_red_euc*: euclidean distance between prompt's and target's TF-IDF vectors projected by UMAP;
- *tfidf_red_cos*: cosine distance between prompt's and target's TF-IDF vectors projected by UMAP;
- *upper_case_word_count_t_over_tp*: the number of upper case words in target divided by their sum in prompt and target;
- *word_density_t_over_tp*: word density in target divided by the sum in prompt and target;
- *punctuation_count_t_over_tp*: the number of punctuation marks in target divided by their sum in prompt and target;
- *char_count_t_over_tp*: the number of characters in target divided by their sum in prompt and target;
- *word_count_t_over_tp*: the number of words in target divided by their sum in prompt and target;
- *tense_tar_per*: the size of the set of tenses in both target and prompt divided by the one in the target only;
- *tense_iou*: the size of the set of tenses in both target and prompt divided by the one in either target or prompt;
- *ent_tar_per*: the size of the set of entities in both target and prompt divided by the one in the target only;
- *ent_iou*: the size of the set of entities in both target and prompt divided by the one in either target or prompt;
- *sbert_umap1_pro*: first component of the 2d UMAP projection of the average vector from the embedding of prompt's sentences;
- *sbert_umap2_pro*: second component of the 2d UMAP projection of the average vector from the embedding of prompt's sentences;
- *sbert_umap1_tar*: first component of the 2d UMAP projection of the vector from the embedding of target's sentence;
- *sbert_umap2_tar*: second component of the 2d UMAP projection of the vector from the embedding of target's sentence;
- *sbert_red_euc*: euclidean distance between 30d UMAP projection of the average vector from the embedding of prompt's sentences and the vector from the embedding of target's sentence;
- *sbert_red_cos*: cosine distance between 30d UMAP projection of the average vector from the embedding of prompt's sentences and the vector from the embedding of target's sentence;
- *emb_cos_mean*: average of the pairwise cosine distances between the prompt's sentence embedding vectors and the target's;
- *emb_cos_max*: maximum of the pairwise cosine distances between the prompt's sentence embedding vectors and the target's;

- *emb_cos_last*: the pairwise cosine distances between the vector embedding of the prompt's last sentence and the target's;
- *emb_cos_min*: minimum of the pairwise cosine distances between the prompt's sentence embedding vectors and the target's.

These features were then passed to a machine learning model that classifies whether the target sentences were coherently following the prompt text, thus tackling the Subtask 1 of the challenge.

The classifier model of choice was LGBMClassifier, a popular machine learning solution that combines computational efficiency with good predictive performances in various problems. The model was imported from the LightGBM gradient boosting framework that uses tree-based learning algorithms [24], and the binary cross-entropy was set as the objective function for the training. Model's hyperparameters were selected by stratified 10-fold cross-validation on shuffled data [25], exhaustively searching the space of hyperparameters (*num_leaves* $\in \{24, 25, 26\}$, *max_depth* $\in \{5, 6, -1\}$, *learning_rate* $\in \{0.009, 0.01, 0.011\}$, *n_estimators* $\in \{185, 190, 195\}$) for the combination with best overall accuracy. This search space of hyperparameters was defined empirically, starting with intervals centred around extreme parameters (*num_leaves* 50, *max_depth* 10, *learning_rate* 0.01, and *n_estimators* 1000). Then, for every new training instance, those intervals were re-centred at the previous best-fitting value if such value stood at the extremity of the interval. Otherwise, the interval was made more narrow. The random seed was set to 42 for reproducibility.

The model's performance was evaluated against the baseline provided by the challenge organizers. See [20] for more details.

3. Results

Training the system on the data available for Subtask 1, the model achieved an accuracy of 0.595 on the test set, improving upon the challenge baseline (0.525) by only 0.07 points. To provide some insights into these performances, the confusion matrix on the training data and the importance of the features are shown in Figure 1 and Figure 2, respectively. Finally, the relevant hyperparameters, resulting from the grid-search cross-validation, are reported in Table 1.

4. Discussion

The features' importance highlights that standard frequency statistics (such as comparisons between prompt and target on TF-IDF vectors and counts of upper case words, tenses, punctuation, words, and characters) are not very informative in predicting coherence for the system employed. Yet "*ent_tar_per*", the proportion of target's entities also present in the prompt's text, was rather important. Semantic information seemed the most relevant, as supported by the many important features extracted by leveraging the sentence embedding model. Of note, "*sbert_umap1_tar*", the projection of the target's sentence embedding, was quite important although not derived from the comparison between prompt and target.

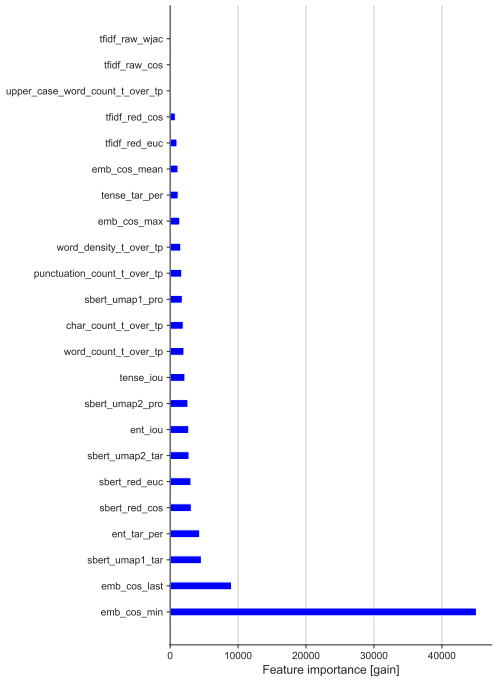


Figure 1: Importance of features from the decision tree-based classifier model.

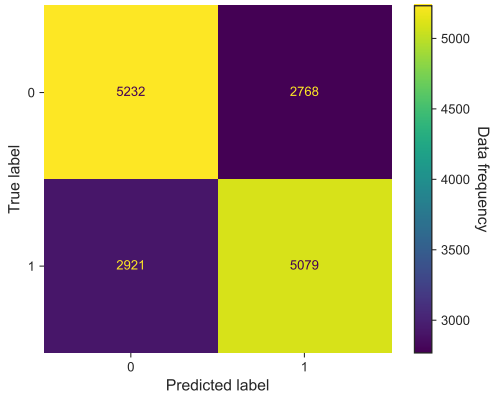


Figure 2: Confusion matrix from the classifier model's predictions on the training data.

Table 1
Best-fitting hyperparameters from the grid-search cross-validation of the classifier model.

Hyperparameter	Value
boosting_type	gbdt
colsample_bytree	1
importance_type	gain
learning_rate	0.01
max_depth	6
min_child_samples	20
min_child_weight	0.001
min_split_gain	0
n_estimators	185
num_leaves	25
objective	binary
random_state	42
metric	accuracy

Finally, the most important features resulted from the aggregation of the pairwise cosine distances between the sentence embedding of the target sentence against those from the prompt sentences. While the average and the maximum of these distances seemed not much important ("*emb_cos_mean*'', "*emb_cos_max*''), the minimum distance and the distance between the target's and prompt's last sentences ("*emb_cos_min*'', "*emb_cos_last*'') were the two most important features. These findings suggest that coherence is elicited by one or a few proximal sentences, while the rest might be of secondary importance.

Our results suggest that including syntactic and discourse-level features might lead to improved performances. Syntactic features, such as part of speech tagging, dependency relationships, or parsing trees, can provide insights into sentence structure and overall grammatical coherence. Moreover, discourse-level features, such as entity co-reference, readability metrics, argumentative structure, discourse markers or topics progression-related features, could assess the flow of ideas in the documents provided. Future work will address the extraction of discourse structure and syntactic features to enable our model to assess a certain text's logical connections, organization and grammatical structure.

The moderate performance improvement might also suggest limitations of standard machine learning models. This could be due to several reasons. First, these models have limited representation capacity of semantic relationships compared to deep learning models, as they lack the sequential modelling needed to represent a text's underlying coherence. Moreover, they lack automatic contextual understanding, focusing more on provided features that might not capture the global context appropriately. Finally, they generally struggle with high-dimensional data. If the number of features is high, as in text coherence, the model could perform sub-optimally, with increased computational complexity. Future improvements on the modelling will include the use of deep learning techniques such as CNNs [11, 12] and Sequence Models [3, 13].

Acknowledgments

This work has been supported by the Horizon Europe VALAWAI project (grant agreement number 101070930). We also wish to thank the Evalita-23 organizers for organizing the task and emphasizing the significance of text coherence measures for the Italian language. The creation and annotation of the corpus have been instrumental in advancing the field of Natural Language Processing for the Italian language and fostering community interest in coherence assessment. Your efforts will enable us to develop AI-driven methods for fostering comprehension assessment applied to both the infosphere and hybrid speech and language practices.

References

[1] Y. Jernite, S. R. Bowman, D. Sontag, Discourse-based objectives for fast unsupervised sentence representation learning, arXiv preprint arXiv:1705.00557 (2017).

[2] Y. Wu, B. Hu, Learning to extract coherent summary via deep reinforcement learning, in: Proceedings of the AAAI conference on artificial intelligence, volume 32, 2018.

[3] J. Li, D. Jurafsky, Neural net models of open-domain discourse coherence, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 198–209. URL: https://aclanthology.org/D17-1019. doi:10.18653/v1/D17-1019.

[4] Z. Lin, H. T. Ng, M.-Y. Kan, Automatically evaluating text coherence using discourse relations, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, pp. 997–1006.

[5] M. Zhang, V. W. Feng, B. Qin, G. Hirst, T. Liu, J. Huang, Encoding world knowledge in the evaluation of local coherence, in: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015, pp. 1087–1096.

[6] G. Albertin, A. Miaschi, D. Brunato, On the role of textual connectives in sentence comprehension: A new dataset for italian., in: CLiC-it, 2021.

[7] R. Barzilay, M. Lapata, Modeling local coherence:

An entity-based approach, Computational Linguistics 34 (2008) 1–34.
[8] M. Elsner, E. Charniak, Extending the entity grid with entity-specific features, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, pp. 125–129.
[9] B. J. Grosz, A. K. Joshi, S. Weinstein, Centering: A framework for modelling the local coherence of discourse, IRCS Technical Reports Series (1995).
[10] M. Lapata, R. Barzilay, et al., Automatic evaluation of text coherence: Models and representations, in: Ijcai, volume 5, 2005, pp. 1085–1090.
[11] D. T. Nguyen, S. Joty, A neural local coherence model, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 1320–1330.
[12] H. C. Moon, T. Mohiuddin, S. Joty, X. Chi, A unified neural coherence model, arXiv preprint arXiv:1909.00349 (2019).
[13] M. Mesgar, M. Strube, A neural local coherence model for text quality assessment, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 4328–4339. URL: https://aclanthology.org/D18-1464. doi:10.18653/v1/D18-1464.
[14] Y. Farag, H. Yannakoudakis, Multi-task learning for coherence modeling, arXiv preprint arXiv:1907.02427 (2019).
[15] A. Beyer, S. Loáiciga, D. Schlangen, Is incoherence surprising? targeted evaluation of coherence prediction from language models, arXiv preprint arXiv:2105.03495 (2021).
[16] Y. Farag, J. Valvoda, H. Yannakoudakis, T. Briscoe, Analyzing neural discourse coherence models, arXiv preprint arXiv:2011.06306 (2020).
[17] A. Lai, J. Tetreault, Discourse coherence in the wild: A dataset, evaluation and methods, arXiv preprint arXiv:1805.04993 (2018).
[18] L. Pishdad, F. Fancellu, R. Zhang, A. Fazly, How coherent are neural models of coherence?, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 6126–6138.
[19] A. Shen, M. Mistica, B. Salehi, H. Li, T. Baldwin, J. Qi, Evaluating document coherence modeling, Transactions of the Association for Computational Linguistics 9 (2021) 621–640.
[20] D. Brunato, D. Colla, F. Dell'Orletta, I. Dini, D. P. Radicioni, A. A. Ravelli, Discotex at evalita 2023: Overview of the assessing discourse coherence in italian texts task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.
[21] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.
[22] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019).
[23] L. McInnes, J. Healy, N. Saul, L. Grossberger, Umap: Uniform manifold approximation and projection, The Journal of Open Source Software 3 (2018) 861.
[24] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T. Liu, Lightgbm: A highly efficient gradient boosting decision tree, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.
[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

Author Index

Almela, Ángela, 79
Altuna, Begoña, 279
Alzetta, Chiara, 130
Apiletti, Daniele, 118
Araque, Oscar, 28

Baralis, Elena, 118
Barbini, Matilde, 144, 326
Basile, Pierpaolo, 299
Basile, Valerio, 11
Basili, Roberto, 11
Benedetto, Irene, 118
Bianco, Selene, 257
Bonanni, Francesca, 150
Bondielli, Alessandro, 213
Bosco, Cristina, 155
Brasolin, Paolo, 150
Brivio, Matteo, 85
Brunato, Dominique, 130, 318
Bucci, Alessandro, 262
Bulla, Luana, 61

Cabrera-Pineda, Hiram, 73
Cageggi, Gioele, 42
Cagliero, Luca, 105, 118, 232
Calo', Gianluca, 36
Can, Fidan, 85
Carman, Mark James, 244
Caselli, Tommaso, 65, 180
Cassotti, Pierluigi, 299
Casula, Camilla, 97, 206
Celli, Fabio, 155
Chakravarthi, Bharathi Raja, 194
Chesi, Cristiano, 144, 326
Cignarella, Alessandra Teresa, 180
Cignoni, Giacomo, 262
Colla, Davide, 318
Croce, Danilo, 11
Cutugno, Francesco, 53
Çöltekin, Çağrı, 85

Damo, Greta, 180
Das, Dipankar, 221
Dascalu, Mihai, 251
De Carolis, Berardina, 36
Dell'Oglio, Pietro, 213
Dell'Orletta, Felice, 130, 318
Di Bonaventura, Chiara, 20
Di Rosa, Emanuele, 42
Dini, Irene, 318
Dinu, Liviu P., 287
Dubossarsky, Haim, 307
D'Amico, Lorenzo, 232

Esuli, Andrea, 226

Franzini, Greta H., 150
Fraser, Alexander, 168
Frenda, Simona, 28
Frey, Jennifer-Carmen, 150

Gafà, Giovanni, 53
Galletti, Martina, 333
Gallipoli, Giuseppe, 105
Gallo, Simone, 112
García-Díaz, José Antonio, 65
García-Sánchez, Francisco, 79
Gatto, Maristella, 299
Giobergia, Flavio, 118, 268
Gonzalo, Julio, 10
Gravino, Pietro, 333
Greco, Salvatore, 105
Gregori, Lorenzo, 313
Grotti, Leonardo, 174

Hangya, Viktor, 168
Hromei, Claudiu D., 11

Ieva, Francesca, 293

Jiang, Shengyi, 138
Jiménez-Zafra, Salud María, 65

Karunakaran, Goutham, 279
Koudounas, Alkis, 118
Kumaresan, Prasanna Kumar, 194

La Quatra, Moreno, 105
Labruna, Tiziano, 112
Lai, Mirko, 1, 3, 155
Lavelli, Alberto, 279
Lenci, Alessandro, 213
Leonardelli, Elisa, 206
Lin, Nankai, 138
Locatelli, Davide, 201
Locatelli, Lorenzo, 201

Loglisci, Corrado, 36
Lopopolo, Olga, 150

Magnini, Bernardo, 279
Marcelloni, Francesco, 213
Massafra, Francesco, 36
Menini, Stefano, 1, 3
Miaschi, Alessio, 130
Micluța-Câmpeanu, Marius, 287
Miranda, Sabino, 73
Monaco, Simone, 118
Mongiovì, Misael, 61
Muti, Arianna, 20

Napolitano, Davide, 232
Nguyen, Dai Ba, 48
Nozza, Debora, 28, 180

Paccosi, Teresa, 274
Palmero Aprosio, Alessio, 274
Pan, Ronghao, 79
Paraschiv, Andrei, 251
Passaro, Lucia C., 213, 299
Patti, Viviana, 28, 155, 180
Pellegrino, Filippo, 150
Periti, Francesco, 307
Pirrone, Roberto, 163, 188
Polignano, Marco, 1, 3
Ponnusamy, Kishore Kumar, 194
Ponnusamy, Rahul, 194
Prevedello, Giulio, 333
Priyadharshini, Ruba, 194
Puccetti, Giovanni, 226

Quick, Patrick, 174

Radicioni, Daniele Paolo, 318
Rajkumar, Charmathi, 194
Ramponi, Alan, 97, 155
Ravelli, Andrea Amelio, 318
Rege Cambrin, Daniele, 105
Ribeiro, Manoel Horta, 238
Rodríguez-García, Miguel Ángel, 91
Russo, Daniel, 65
Russo, Giuseppe, 238
Russo, Valentina, 1, 3

Sabbatini, Marco, 213
Sagae, Kenji, 130
Salogni, Ilaria, 124
Salusso, Daniela, 257
Sarkar, Sandip, 221
Scotti, Vincenzo, 244
Siciliani, Lucia, 299
Siragusa, Irene, 163, 188
Speranza, Manuela, 279
Spina, Stefania, 150
Sprugnoli, Rachele, 1, 3, 28
Stemle, Egon W., 150
Stoehr, Niklas, 238
Stranisci, Marco Antonio, 20
Sánchez-Gutiérrez, Claudia H., 130

Tebaldini, Martina, 150
Thin, Dang Van, 48
Tonelli, Sara, 155
Torri, Vittorio, 293
Tudu, Nripen, 221
Téllez, Eric Sadit, 73

Uboldi, Asia, 42
Ureña-López, L. Alfonso, 65
Uyen, Nguyen Ngoc Phuong, 48

Vaiani, Lorenzo, 232
Valencia-García, Rafael, 65
Venturi, Giulia, 1, 3, 130
Venuti, Marco, 53
Vitali, Michael, 244

Wu, Hongyan, 138

Xiao, Lixian, 138

Zanoli, Emma, 144, 326
Zanoli, Roberto, 279