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

Francielle Vargas
USP
francielleavargas@usp.br

Thiago Pardo
USP
taspardo@icmc.usp.br

Fabrício Benevenuto
UFMG
fabricio@dcc.ufmg.br

ABSTRACT

Although Natural Language Processing (NLP) has traditionally relied on inherently interpretable "white-box" techniques, such as rule-based algorithms, decision trees, hidden Markov models, and logistic regression, the adoption of Large Language Models (LLMs) and language embeddings (often considered "black-box") has significantly reduced interpretability. This lack of transparency introduces considerable risks, including biases, which have become a major concern in the field of Artificial Intelligence (AI). This Ph.D. thesis addresses these critical gaps by proposing new resources that ensure interpretability and fairness in NLP models for automated fact-checking and hate speech detection tasks. Specifically, we introduce five benchmark datasets (HateBR, HateBRXplain, HausaHate, MOL, and FactNews), four novel post-hoc and self-explaining methods (SELFAR, SSA, B+M and SRA), and one web platform (NoHateBrazil) designed to improve the interpretability and fairness of hate speech detection. The proposed models outperform the existing baselines for Portuguese and Hausa, both underrepresented languages. This research contributes to ongoing discussions on responsible and explainable AI, bridging the gap between model performance and interpretability to achieve positive real-world social impact. Finally, this thesis has had a significant impact both nationally and internationally, receiving citations from prestigious universities and research institutes abroad, inspiring new M.Sc. and Ph.D. in Brazil, and being recognized with multiple awards, including the Google LARA, the Maria Carolina Monard AI Award, and as a finalist at the Brazilian Computer Society Thesis and Dissertation Award.

KEYWORDS

natural language processing, explanability and interpretability, social networks and social media, misinformation, hate speech, and online toxicity, fairness, responsible ai.

1 INTRODUCTION

According to the United Nations (UN) and a growing body of literature, misinformation exacerbates hate speech against individuals or groups based on their social identity [7, 18]. Hate speech often reinforces social divisions by categorizing individuals into in-groups and out-groups, concepts rooted in social identity theory [9]. Furthermore, hate speech can deliberately spread falsehoods about out-groups to deepen social polarization and strengthen support for radical ideological positions [3].

To address the harmful cycle, a wide range of NLP methods have been proposed for automated fact-checking and hate speech detection. Nevertheless, although NLP has historically relied on inherently interpretable methods, the rise adoption of LLMs (also known as "black-box" or "opaque box") and language embeddings have significantly reduced the interpretability of NLP models [10]. Consequently, most existing fact-checking and hate speech detection models do not provide rationales for their predictions. This lack of transparency introduces significant risks, including biases, which have become a major concern in the field of AI [6].

Regarding automated fact checking and hate speech detection tasks, this lack of transparency raises serious concerns regarding the robustness, reliability, and fairness of models [1, 2]. In response to these critical issues, this thesis acknowledges the detrimental impact of insufficient transparency, particularly in applications aimed at combating misinformation and hate speech, both of which are essential to maintaining a fair and democratic society. Specifically, we introduced five benchmark datasets (HateBR [13], HateBRXplain [8], HausaHate [14], FactNews [15], and MOL [12]) and developed four novel post-hoc and self-explaining methods (SELFAR [17], SSA [11], B+M [16] and SRA) to ensure that both the data and the models used to tackle misinformation and hate speech are explainable and mitigate bias. Notably, HateBR and B+M outperformed existing baselines in Portuguese, while SRA surpassed current baselines in English, providing more faithful explanations and achieving better bias mitigation scores. Overall, this thesis advances research in Explainable AI and fairness within automated fact-checking and hate speech detection tasks, while also fostering broader discussions on interpretability and fairness in the fields of NLP and Machine Learning (ML).

2 RESEARCH METHODOLOGY

The research follows a methodology comprising: (i) development of benchmark datasets with expert annotations and human rationales; (ii) implementation of explainable machine learning architectures such as post-hoc and self-explaining; (iii) evaluation using both standard NLP metrics and explainability, and fairness assessments; (iv) comparative analysis against state-of-the-art models to validate improvements in robustness, interpretability, and fairness.

3 AWARDS & GRANTS

This work has been widely recognized with multiple awards and grants, including the Maria Carolina Monard Award in AI (2025), finalist for the SBC Thesis Award (2025), Google LARA Award fellowship (2024), and ACL travel grants (EMNLP, NAACL 2024).

In: VII Concurso de Teses e Dissertações (CTD 2025). Anais Estendidos do XXXI Simpósio Brasileiro de Sistemas Multimídia e Web (CTD'2025). Rio de Janeiro/RJ, Brasil. Porto Alegre: Brazilian Computer Society, 2025.

ISSN 2596-1683

4 RESEARCH IMPACT

Several prestigious research institutions, including universities and industry centers, have been significantly influenced by the provided resources of the thesis, as reflected in a substantial number of citations. Regarding **international impact**, institutions including Microsoft Research, Carnegie Mellon University, Rochester Institute of Technology, Harvard University, University of Maryland, University of Turin, Technical University of Munich, University of Bonn, University Institute of Lisbon, National University of Singapore, and Vrije Universiteit Amsterdam have cited or used our resources. Regarding **national impact**, universities such as USP, UFF, UFCG, UDESC, UFOP, and UFMG have proposed new Ph.D. and MSc theses focused on the study of hate speech and fake news, applying our resources. Furthermore, the research conducted in this thesis led to invitations to serve as a visiting researcher at the University of Southern California (USC) in the USA, and to visit and speak at the GESIS - Leibniz Institute for Social Science in Germany. I was also invited to serve on the organizing committee for the International Conference on Web and Social Media (ICWSM) in 2021, 2022, and 2023, as well as the ACL Workshop on Online Abuse and Harms (WOAH) and the IJCNN Special Session Explainable Deep Neural Networks for Responsible AI (DeepXplain) in 2025. Lastly, I have been an active participant in the program committee for several top-tier NLP conferences and workshops, including ACL, EMNLP, NAACL, LREC, COLING, WOAH, FEVER, and CODI.

5 CONTRIBUTIONS

The proposed datasets, methods and system significantly advance research in hate speech, fact-checking, and explainable AI, mainly for low-resource languages. The impact of this research contributing directly to foundational topics in Multimedia, Hypermedia and Web, including: (i) **benchmarking for future research**: The datasets created in this thesis (e.g. HateBR, HateBRXplain, HausaHate, MOL, and FactNews) establish new benchmarks in automated fact-checking and hate speech classification for Portuguese and Hausa languages, enabling future studies to compare and refine models; (ii) **advancements in explainable AI (XAI)**: This thesis pioneers novel methods that enhance interpretability in NLP (e.g., B+M, SRA and SELFAR). By introducing new explainability mechanisms, it sets a new standard for developing ethical and transparent AI systems; (iii) **bias mitigation in NLP**: This research provides innovative solutions to mitigate biases in NLP models (e.g., SSA method) influencing subfields such as fairness-aware machine learning and computational social science; (iv) **low-resource language processing**: By developing datasets and models for Portuguese and other underrepresented languages, such as Hausa, an African indigenous language, this work expands the scope of NLP beyond English, making NLP technologies more inclusive and globally applicable; (v) **positive social impact in real-world applications**: The resources introduced in this thesis have already been adopted by prestigious research institutions worldwide, including Microsoft Research, which has used the HateBR dataset to train two LLMs: CultureLLM [4] and CulturePark [5], demonstrating its clear relevance and practical applicability. Finally, the methods and systems developed in this research may be applied for patents and registered systems with copyrights.

6 PUBLICATIONS

In total, 15 (fifteen) papers were published in top-tier international NLP and AI conferences and journals, including 10 (ten) in Qualis A1 and 5 (five) in Qualis A3 venues: https://franciellevargas.github.io/.

REFERENCES

[1] Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate Speech Classifiers Learn Normative Social Stereotypes. *Transactions of the Association for Computational Linguistics* 11 (2023), 300–319.

[2] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of the 3rd Workshop on Abusive Language Online*. Florence, Italy, 25–35.

[3] Michael Hameleers, Toni Van der, and Rens Vliegenthart. 2022. Civilized truths, hateful lies? Incivility and hate speech in false information – evidence from fact-checked statements in the US. *Information, Communication & Society* 25, 11 (2022), 1596–1613.

[4] Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. CultureLLM: Incorporating Cultural Differences into Large Language Models. In *Advances in Neural Information Processing Systems*, Vol. 37. 84799–84838.

[5] Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2025. CulturePark: boosting cross-cultural understanding in large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '24)*. Red Hook, NY, USA, Article 2082, 34 pages.

[6] Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On Measuring Social Biases in Sentence Encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Minneapolis, Minnesota, 622–628.

[7] Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation* 55, 3 (2021), 477–523.

[8] Isadora Salles, Francielle Vargas, and Fabrício Benevenuto. 2025. HateBRXplain: A Benchmark Dataset with Human-Annotated Rationales for Explainable Hate Speech Detection in Brazilian Portuguese. In *Proceedings of the 31st International Conference on Computational Linguistics*. Abu Dhabi, UAE, 6659–6669.

[9] Henri Tajfel. 1979. An integrative theory of intergroup conflict. *The social psychology of intergroup relations/Brooks/Cole* (1979).

[10] Yulia Tsvetkov, Vinodkumar Prabhakaran, and Rob Voigt. 2019. Socially Responsible Natural Language Processing. In *Companion Proceedings of The 2019 World Wide Web Conference* (San Francisco, USA) *(WWW '19)*. New York, USA, 1326.

[11] Francielle Vargas, Isabelle Carvalho, Ali Hürriyetoğlu, Thiago Pardo, and Fabrício Benevenuto. 2023. Socially Responsible Hate Speech Detection: Can Classifiers Reflect Social Stereotypes?. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*. Varna, Bulgaria, 1187–1196.

[12] Francielle Vargas, Isabelle Carvalho, Thiago Pardo, and Fabricio Benevenuto. 2024. Context-Aware and Expert Data Resources for Brazilian Portuguese Hate Speech Detection. *Natural Language Processing* (2024), 1–22.

[13] Francielle Vargas, Isabelle Carvalho, Fabiana Rodrigues de Góes, Thiago Pardo, and Fabrício Benevenuto. 2022. HateBR: A Large Expert Annotated Corpus of Brazilian Instagram Comments for Offensive Language and Hate Speech Detection. In *Proceedings of the 13th Language Resources and Evaluation Conference*. Marseille, France, 7174–7183.

[14] Francielle Vargas, Samuel Guimarães, Shamsuddeen Hassan Muhammad, Diego Alves, Ibrahim Said Ahmad, Idris Abdulmumin, Diallo Mohamed, Thiago Pardo, and Fabrício Benevenuto. 2024. HausaHate: An Expert Annotated Corpus for Hausa Hate Speech Detection. In *Proceedings of the 8th Workshop on Online Abuse and Harms*. Mexico City, Mexico, 52–58.

[15] Francielle Vargas, Kokil Jaidka, Thiago Pardo, and Fabrício Benevenuto. 2023. Predicting Sentence-Level Factuality of News and Bias of Media Outlets. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*. Varna, Bulgaria, 1197–1206.

[16] Francielle Vargas, Fabiana Rodrigues de Góes, Isabelle Carvalho, Fabrício Benevenuto, and Thiago Pardo. 2021. Contextual-Lexicon Approach for Abusive Language Detection. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*. Held Online, 1438–1447.

[17] Francielle Vargas, Isadora Salles, Diego Alves, Ameeta Agrawal, Thiago A. S. Pardo, and Fabrício Benevenuto. 2024. Improving Explainable Fact-Checking via Sentence-Level Factual Reasoning. In *Proceedings of the 7th Fact Extraction and VERification Workshop*. Miami, USA, 192–204.

[18] Claire Wardle. 2024. *A Conceptual Analysis of the Overlaps and Differences between Hate Speech, Misinformation and Disinformation*. Department of Peace Operations (DPO). Office of the Special Adviser on the Prevention of Genocide (OSAPG). United Nations.