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

Francielle Vargas[1], **Thiago Pardo (Advisor)**[1], **Fabrício Benevenuto (Co-Advisor)**[2]

[1]University of São Paulo

[2]Federal University of Minas Gerais

francielleavargas@usp.br, taspardo@icmc.usp.br, fabricio@dcc.ufmg.br

Abstract. *Misinformation and hate speech form a socially harmful cycle. Research shows that misinformation can amplify hate speech targeting social identity groups and reinforce harmful stereotypes. To combat this cycle, a wide range of Natural Language Processing (NLP) methods have been proposed. Nevertheless, while NLP has historically relied on inherently explainable "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 AI. This Ph.D. thesis addresses these critical gaps by proposing new resources that ensure explainability and bias mitigation in NLP models for these tasks. Specifically, it introduces five benchmark datasets (HateBR, HateBRXplain, HausaHate, MOL, and FactNews), three novel methods (SELFAR, SSA, and B+M), and one web system (NoHateBrazil) designed to improve the explainability and fairness of automated fact-checking and hate speech detection. The proposed models outperform 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 for real-world applications. Finally, it has had a significant impact both nationally and internationally, receiving citations from prestigious universities and research institutes abroad, and inspiring new M.Sc. and Ph.D. projects in Brazil.*

Resumo. *Desinformação e discurso de ódio formam um ciclo socialmente prejudicial. Pesquisas mostram que a desinformação pode amplificar o discurso de ódio contra grupos badeado na sua identidade social e reforçar estereótipos prejudiciais. Para combater esse ciclo, uma ampla variedade de métodos de Processamento de Linguagem Natural (PLN) tem sido proposto. No entanto, embora o PLN tenha historicamente se baseado em técnicas inerentemente explicáveis, conhecidas como "caixa branca", como algoritmos baseados em regras, árvores de decisão, modelos ocultos de Markov e regressão logística, a adoção de Grandes Modelos de Linguagem (LLMs) e embeddings de linguagem (frequentemente considerados "caixa preta"), reduziu significativamente a interpretabilidade. Essa falta de transparência introduz riscos consideráveis, incluindo vieses, que se tornaram uma preocupação importante na IA. Esta tese de doutorado aborda essas lacunas críticas propondo novos recursos que garantem explicabilidade e mitigação de vieses em modelos de PLN para essas*

tarefas. Especificamente, essa tese introduz cinco datasets benchmark (HateBR, HateBRXplain, HausaHate, MOL e FactNews), três métodos inovadores (SELFAR, SSA e B+M) e um sistema web (NoHateBrazil) projetados para melhorar a explicabilidade e a justiça da verificação automática de fatos e da detecção de discurso de ódio. Os modelos propostos superam os baselines para o português e o hausa, ambos idiomas sub-representados. Esta pesquisa contribui para as discussões em curso sobre IA responsável e explicável, preenchendo a lacuna entre desempenho dos modelos e interpretabilidade para aplicações no mundo real. Por fim, os resultados obtidos nessa tese tiveram um impacto significativo tanto nacional quanto internacionalmente, recebendo citações de universidades e institutos de pesquisa de prestígio no exterior e inspirando novos projetos de mestrado e doutorado no Brasil.

1. Introduction

The proliferation of misinformation and hate speech has become a pressing global issue. These phenomena not only distort public opinion but also exacerbate social tensions, often leading to real-world harm. Although the literature has often presented vague definitions of misinformation, disinformation, and hate speech, according to the United Nations (UN) (Wardle 2024), they represent distinct forms of "harmful speech" that, despite their differences, share common characteristics and can have significant long-term impacts. Hate speech, for example, specifically targets individuals or groups based on their social identity. Similarly, disinformation and misinformation can be used to damage reputations or as strategies to spread hate; however, they may also target individuals and groups based on non-identity-related characteristics, such as occupation. The dissemination of fake news can also serve as a tool for discrimination against individuals or groups, potentially reaching the threshold of incitement. In other words, when hate speech is amplified by misinformation, its reach expands, thereby increasing the likelihood of harm (Wardle 2024). An increasing body of literature supports the argument that misinformation exacerbates hate speech, including phenomena such as partyism* (Pennycook and Rand 2018; Poletto et al. 2021; Marwick and Lewis 2017). Hate speech often constructs social divisions by categorizing individuals into in-groups and out-groups†. Out-groups are often described using negative stereotypes and framed with pejorative associations, employing hostile and demeaning language to portray "the other."Furthermore, hate speech can deliberately spread falsehoods about out-groups to deepen social polarization and strengthen support for radical right-wing ideological positions (Hameleers et al. 2022).

To address the harmful cycle of fake news dissemination and its role in triggering hate speech against social groups, a wide range of natural language processing (NLP) methods have been proposed for fact-checking and hate speech detection. Nevertheless, although NLP has historically relied on inherently explainable methods, often

*Partyism refers to extreme hostility toward political affiliations, influencing judgments and behaviors beyond the political sphere (Sunstein 2015; Westwood et al. 2018).

†The concepts of in-group and out-group are rooted in social identity theory (Tajfel 1979). This theory posits that individuals derive part of their self-concept from their membership in social groups, influencing cognitive processes and intergroup behaviors, particularly those related to prejudice, bias, and discrimination.

termed 'white-box' techniques, such as rule-based algorithms, decision trees, hidden Markov models and logistic regressions, the rise adoption of large language models (LLMs) (also known as "black-box"*) and language embeddings have significantly reduced the interpretability of NLP models (Tsvetkov et al. 2019). Consequently, most existing fact-checking and hate speech detection models that rely on these black-box approaches fail to provide rationales for their predictions. This lack of transparency introduces significant risks, including biases, which have become a major concern in the field (May et al. 2019). For example, biases in training data, resulting from prejudiced labels or under- or over-sampling, lead to models with unwanted biases (Stryker 2024).

Hate speech detection technologies often inherit biases from their training data (Davidson et al. 2019), which may reflect subjective human annotations (Al Kuwatly et al. 2020; Sap et al. 2022). These biases can reinforce social discrimination—such as racial and gender biases—especially when deployed at scale (Davani et al. 2023; Chuang et al. 2021; Sap et al. 2019; Davidson et al. 2019). Furthermore, a notable flaw in neural hate speech classifiers is their tendency to overemphasize group identifiers, such as "Muslim," "gay," and "black" (Dixon et al. 2018). These terms are not inherently offensive, but may only constitute hate speech within specific contexts. Table 1 shows this issue, showing two documents classified as hate speech by a fine-tuned BERT classifier. Note that while the second document from Gab contains explicit hate speech, the first document, extracted from the New York Times, does not. However, the fine-tuned BERT classifier incorrectly classified both as hate speech, due to the presence of group identifiers such as "black' and "Africans."

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

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

In similar settings, the issue of bias in fact-checking has been extensively analyzed in the literature (Park et al. 2021; Soprano et al. 2024). Biases (e.g., media bias, political bias, etc.) in automated fact-checking may be also introduced during data training. Consequently, fact-checking models tend to rely on these biases without fully learning the underlying task. Instead, they often learn misleading correlations between news patterns and veracity labels as simplifications, rather than integrating the information to reason effectively (Wu et al. 2022). As a result, these models may not only fail when applied to real-life situations, where news patterns vary widely, but they can also undermine public trust and exacerbate political polarization (Kuzmin et al. 2020). For example, prior studies assessing the performance of human fact-checkers have reported conflicting findings (Amazeen 2015) and identified significant inconsistencies among major

*Black-box techniques generate predictions based on input data, yet their decision-making processes remain opaque to users (Garg et al. 2021). This is is also described as an "opaque box."

fact-checking organizations such as PolitiFact[*], The Fact Checker[†], and FactCheck.org[‡] in their evaluations of statements on topics like climate change, racism, and national debt (Marietta et al. 2015). In addition, only 10% of statements were fact-checked by both organizations in their study, with agreement primarily observed for statements classified as clearly true or false, but significantly lower agreement for ambiguous claims that highlight the inherent subjectivity in the manual assignment of veracity ratings, raising concerns about potential biases such as selective claim verification and inconsistencies in evaluation criteria (Nieminen and Rapeli 2019).

Table 2 presents examples of manually fact-checked claims and their assigned veracity ratings by PolitiFact. Notice that the first claim[§] was rated as "true." However, the journalist reviewing this statement noted that it does not account for differences in occupations between the two populations; thus, had this context been considered, the claim could have been deemed misleading. Moreover, another sentence in the same tweet was not fact-checked. Similarly, the second claim[¶] received a "mostly true" rating, with the justification that while the statement itself is accurate, it lacks crucial context. In both cases, the assigned veracity ratings may be imprecise, leading to ambiguous justifications.

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

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

Therefore, the inability of NLP models to provide rationales for their decisions remains a significant barrier to their broader adoption (Gongane et al. 2024). In the context of automated fact-checking and hate speech detection, this lack of transparency raises serious ethical concerns regarding model reliability and fairness. In response to these critical issues, this thesis 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. These challenges highlight a pressing and timely research opportunity. To address them, this thesis proposed a study focused on the development of socially responsible and explainable technologies for fact-checking and hate speech detection. As part of this effort, we introduced five benchmark datasets (HateBR, HateBRXplain, HausaHate, FactNews, and MOL) and developed three novel computational methods (SELFAR, SSA, and B+M) to ensure that both data and models used to tackle misinformation and hate speech are explainable and socially aligned. Notably, HateBR and B+M outperformed existing baselines for the Portuguese language. Overall, the contributions of this thesis aim to advance research in automated

[*] https://www.politifact.com/

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

[‡] https://www.factcheck.org/

[§] https://www.politifact.com/factchecks/2022/dec/16/tammy-baldwin/yes-wage-gap-does-have-big-impact-latina-workers/?cid=twitter_PolitiFactWisc

[¶] https://www.politifact.com/factchecks/2023/mar/29/william-barclay/syracuse-proposal-would-pay-gang-members-stay-away/

fact-checking and hate speech detection while contributing to the ongoing discussion on explainability, interpretability, and fairness in natural language processing and machine learning.

2. Research Impact

As a result of the research presented in this thesis, a range of national and international projects have been initiated, leveraging the proposed methods and dataset benchmarks. Specifically, various research institutions, including universities and industry partners, have been significantly influenced by the provided resources, as reflected in a substantial number of citations. For example, in the **international context**, institutions such as 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 utilized the HateBR and FactNews datasets, the Multilingual Offensive Lexicon (MOL), and the B+M method in their research endeavors. In the **national context**, universities such as Fluminense Federal University (UFF), Federal University of Campina Grande (UFCG), State University of Santa Catarina (UDESC), Federal University of Ouro Preto (UFOP), and Federal University of Minas Gerais (UFMG) have proposed Ph.D. and MSc theses focused on the study of hate speech, utilizing our data resources, including HateBR, HateBRXplain, and the Multilingual Offensive Lexicon. Furthermore, the research conducted in this thesis led to an invitation to serve as a visiting researcher at the University of Southern California (USC) in the USA, and to speak at the GESIS - Leibniz Institute for Social Science in Germany. I also had the opportunity to serve on the organizing committee for the International Conference on Web and Social Media (ICWSM) in 2021, 2022, and 2023. More recently, I was invited to join the organizing team for the international Workshop on Online Abuse and Harms (WOAH), the leading workshop in Natural Language Processing focused on hate speech. Additionally, I have been an active participant in the program committee for several prominent international NLP and Computational Social Science conferences and workshops, including ACL, EMNLP, NAACL, LREC, COLING, CODI, WOAH, and FEVER. Lastly, I contributed to the Southern California Natural Language Processing Symposium (SoCal NLP 2022) and served as a lead co-organizer of the Explainable Deep Neural Networks for Responsible AI (DeepXplain) special session at the International Joint Conference on Neural Networks (IJCNN 2025).

3. Research Problem and Motivation

Traditional NLP models for automated fact-checking and hate speech detection have relied on inherently explainable techniques such as rule-based systems and decision trees. However, the advent of deep learning and Large Language Models (LLMs) has significantly reduced model transparency. The inability of these models to provide rationales (i.e., justifications) for their decisions introduces significant risks, including the potential to perpetuate biases that disproportionately harm marginalized communities or exacerbate political polarization. This thesis aims to bridge this gap by developing methods that improve the interpretability of NLP models while ensuring their ethical and social responsibility.

4. Objectives

The main objectives of this thesis are:

1. To evaluate the risks and biases in black-box NLP models for fact-checking and hate speech detection.
2. To create new benchmark datasets that enhance model explainability and mitigate biases.
3. To develop novel methods that integrate explainability into NLP models, maintain high predictive performance, and ensure fairness and bias mitigation.
4. To support low-resource languages by providing robust datasets and methods.

5. Contributions

5.1. Benchmark Datasets

- **HateBR** (Vargas et al. 2022) and **HateBRXplain** (Salles et al. 2025): The first large-scale expert-annotated datasets for hate speech detection in Brazilian Portuguese, including human-annotated rationales for explainability.
- **HausaHate** (Vargas et al. 2024b): A benchmark dataset for hate speech detection in Hausa, an indigenous African language primarily spoken in Nigeria.
- **MOL** (Vargas et al. 2024a): The first multilingual offensive lexicon, consisting of 1,000 explicit and implicit terms and expressions with pejorative connotations, annotated with contextual information. The lexicon also includes native-speaker translations and cultural adaptations in English, Spanish, French, German, and Turkish.
- **FactNews** (Vargas et al. 2023c): A benchmark dataset for sentence-level news source reliability estimation in Portuguese news.

5.2. Automated Methods

- **Sentence-Level Factual Reasoning (SELFAR)** (Vargas et al. 2024c): The first study for explainable automated fact-checking in Portuguese. Specifically, the SELFAR relies on fact extraction and verification by predicting the news source reliability and factuality (veracity) of news articles or claims at the sentence level, generating post-hoc explanations using SHAP/LIME and zero-shot prompts.
- **Social Stereotype Bias (SSA)** (Vargas et al. 2023a): SSA is a counterfactual explanation method for assessing social bias in hate speech classifiers through the analysis of stereotypes and counter-stereotypes.
- **B+M** (Vargas et al. 2021): The B+M method consists of a contextualized bag-of-words (BoW) model enhanced with feature saliency for explainable hate speech detection. Our approach adds an additional layer of significance to features from the MOL lexicon by incorporating their contextual labels (e.g., context-dependent vs. context-independent).

5.3. A Benchmark System

- **NoHateBrazil** (Vargas et al. 2023b): A web system for Brazilian Portuguese text offensiveness analysis leveraging a self-explaining rule-based algorithm. Link: `http://143.107.183.175:14581/`

6. Research Methodology

The research follows a data-driven methodology comprising:

- Development of benchmark datasets with expert annotations and rationales.
- Implementation of explainable machine learning techniques, including feature attribution, as well as post-hoc and self-explaining architectures.
- Evaluation using both standard NLP metrics and explainability assessments.
- Comparative analysis against state-of-the-art models to validate improvements in accountability and fairness.

7. Results and Impact on Computer Science

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 Computer Science, including:

- **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.
- **Advancements in Explainable AI (XAI)**: This thesis pioneers novel methods that enhance interpretability in NLP. By introducing new explainability mechanisms, it sets a new standard for developing ethical and transparent AI systems.
- **Bias Mitigation in NLP**: This research provides innovative solutions to mitigate biases in NLP models, influencing subfields such as fairness-aware machine learning and computational social science.
- **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-centric approaches, making NLP technologies more inclusive and globally applicable.
- **Impactful Results in Real-World Applications**: The methodologies introduced in this thesis have already been adopted by prestigious research institutions worldwide, including Microsoft Research, Carnegie Mellon University, and the University of Maryland, demonstrating their clear relevance and practical applicability. Finally, the methods and systems developed in this research may be applied for patents and registered systems with copyrights.

8. Conclusion

This thesis highlights the necessity of integrating explainability into NLP models to ensure the ethical deployment of AI, focusing on critical tasks such as automated fact-checking and hate speech detection. By proposing benchmark datasets and novel methods and systems, it lays the foundation for more transparent and socially responsible AI solutions in misinformation and hate speech detection. Future research will explore the generalization of these techniques to other languages, tasks, and deep learning architectures, further advancing the goal of responsible and explainable AI.

Acknowledgements

This project was partially funded by Google, FAPESP, CNPq, FAPEMIG, CAPES, and the Ministry of Science, Technology and Innovation, with resources of Law N. 8,248, of October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published as Residence in TIC 13, DOU 01245.010222/2022-44.

Referências

[Al Kuwatly et al. 2020] Al Kuwatly, H., Wich, M., and Groh, G. (2020). Identifying and measuring annotator bias based on annotators' demographic characteristics. In Akiwowo, S., Vidgen, B., Prabhakaran, V., and Waseem, Z., editors, *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, Held Online.

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

[Chuang et al. 2021] Chuang, Y.-S., Gao, M., Luo, H., Glass, J., Lee, H.-y., Chen, Y.-N., and Li, S.-W. (2021). Mitigating biases in toxic language detection through invariant rationalization. In *Proceedings of the 5th Workshop on Online Abuse and Harms*, pages 114–120, Held Online.

[Davani et al. 2023] Davani, A. M., Atari, M., Kennedy, B., and Dehghani, M. (2023). Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319.

[Davidson et al. 2019] Davidson, T., Bhattacharya, D., and Weber, I. (2019). Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the 3rd Workshop on Abusive Language Online*, pages 25–35, Florence, Italy.

[Dixon et al. 2018] Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73, New York, USA.

[Garg et al. 2021] Garg, P., Chakravarthy, A. S., Mandal, M., Narang, P., Chamola, V., and Guizani, M. (2021). Isdnet: Ai-enabled instance segmentation of aerial scenes for smart cities. *ACM Transactions on Internet Technology (TOIT)*, 21(3):1–18.

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

[Hameleers et al. 2022] Hameleers, M., Van der, T., and Vliegenthart, R. (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):1596–1613.

[Kennedy et al. 2020] Kennedy, B., Jin, X., Mostafazadeh Davani, A., Dehghani, M., and Ren, X. (2020). Contextualizing hate speech classifiers with post-hoc explanation. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442, Held Online.

[Kuzmin et al. 2020] Kuzmin, G., Larionov, D., Pisarevskaya, D., and Smirnov, I. (2020). Fake news detection for the Russian language. In *Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media*, pages 45–57, Barcelona, Spain.

[Marietta et al. 2015] Marietta, M., Barker, D. C., and Bowser, T. (2015). Fact-checking polarized politics: Does the fact-check industry provide consistent guidance on disputed realities? *The Forum*, 13(4):577–596.

[Marwick and Lewis 2017] Marwick, A. E. and Lewis, B. (2017). Media manipulation and disinformation online. *Data and Society Research Institute*, pages 1 – 104.

[May et al. 2019] May, C., Wang, A., Bordia, S., Bowman, S. R., and Rudinger, R. (2019). On measuring social biases in sentence encoders. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 622–628, Minneapolis, Minnesota.

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

[Park et al. 2021] Park, S., Park, J. Y., Kang, J.-h., and Cha, M. (2021). The presence of unexpected biases in online fact-checking. *Harvard Kennedy School Misinformation Review*, 2(1).

[Pennycook and Rand 2018] Pennycook, G. and Rand, D. G. (2018). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, (188):39–50.

[Poletto et al. 2021] Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., and Patti, V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(3):477–523.

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

[Sap et al. 2019] Sap, M., Card, D., Gabriel, S., Choi, Y., and Smith, N. A. (2019). The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy.

[Sap et al. 2022] Sap, M., Swayamdipta, S., Vianna, L., Zhou, X., Choi, Y., and Smith, N. A. (2022). Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V., editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States.

[Soprano et al. 2024] Soprano, M., Roitero, K., La Barbera, D., Ceolin, D., Spina, D., Demartini, G., and Mizzaro, S. (2024). Cognitive biases in fact-checking and their countermeasures: A review. *Inf. Process. Manage.*, 61(3).

[Stryker 2024] Stryker, C. S. (2024). What is responsible ai? *International Business Machines (IBM)*.

[Sunstein 2015] Sunstein, C. R. (2015). Partysm. pages 1–19.

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

[Tsvetkov et al. 2019] Tsvetkov, Y., Prabhakaran, V., and Voigt, R. (2019). Socially responsible natural language processing. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 1326, New York, USA.

[Vargas et al. 2023a] Vargas, F., Carvalho, I., Hürriyetoğlu, A., Pardo, T., and Benevenuto, F. (2023a). Socially responsible hate speech detection: Can classifiers reflect social stereotypes? In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1187–1196, Varna, Bulgaria.

[Vargas et al. 2024a] Vargas, F., Carvalho, I., Pardo, T., and Benevenuto, F. (2024a). Context-aware and expert data resources for brazilian portuguese hate speech detection. *Natural Language Processing*, pages 1–22.

[Vargas et al. 2022] Vargas, F., Carvalho, I., Rodrigues de Góes, F., Pardo, T., and Benevenuto, F. (2022). HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 7174–7183, Marseille, France.

[Vargas et al. 2023b] Vargas, F., Carvalho, I., Schmeisser-Nieto, W., Benevenuto, F., and Pardo, T. (2023b). NoHateBrazil: A Brazilian Portuguese text offensiveness analysis system. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1180–1186, Varna, Bulgaria.

[Vargas et al. 2024b] Vargas, F., Guimarães, S., Muhammad, S. H., Alves, D., Ahmad, I. S., Abdulmumin, I., Mohamed, D., Pardo, T., and Benevenuto, F. (2024b). HausaHate: An expert annotated corpus for Hausa hate speech detection. In *Proceedings of the 8th Workshop on Online Abuse and Harms*, pages 52–58, Mexico City, Mexico.

[Vargas et al. 2023c] Vargas, F., Jaidka, K., Pardo, T., and Benevenuto, F. (2023c). Predicting sentence-level factuality of news and bias of media outlets. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1197–1206, Varna, Bulgaria.

[Vargas et al. 2021] Vargas, F., Rodrigues de Góes, F., Carvalho, I., Benevenuto, F., and Pardo, T. (2021). Contextual-lexicon approach for abusive language detection. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 1438–1447, Held Online.

[Vargas et al. 2024c] Vargas, F., Salles, I., Alves, D., Agrawal, A., Pardo, T. A. S., and Benevenuto, F. (2024c). Improving explainable fact-checking via sentence-level factual reasoning. In *Proceedings of the 7th Fact Extraction and VERification Workshop*, pages 192–204, Miami, USA.

[Wardle 2024] Wardle, C. (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.

[Westwood et al. 2018] Westwood, S. J., Iyengar, S., Walgrave, S., Leonisio, R., Miller, L., and Strijbis, O. (2018). The tie that divides: Cross-national evidence of the primacy of partyism. *European Journal of Political Research*, 57:333–354.

[Wu et al. 2022] Wu, J., Liu, Q., Xu, W., and Wu, S. (2022). Bias mitigation for evidence-aware fake news detection by causal intervention. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2308–2313, New York, USA.