

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

Francielle Vargas¹, Thiago A. S. Pardo², Fabricio Benevenuto³

¹University of Chile,

²University of São Paulo

³Federal University of Minas Gerais

francielle@uchile.cl, taspardo@icmc.usp.br, fabricio@dcc.ufmg.br

Abstract. *This Ph.D. dissertation advances the state-of-the-art in AI - Natural Language Processing (NLP) by addressing two critical challenges: the scarcity of high-quality benchmark resources for underrepresented languages in the Global South and the limited interpretability of modern language models for socially relevant tasks. The research focuses on hate speech detection and automated fact-checking, proposing novel datasets, explainable methods, and evaluation frameworks to improve the transparency, fairness and reliability of NLP systems. The dissertation introduces seven benchmark datasets for Brazilian Portuguese and the Hausa language: HateBR, HateBRXplain, HateBRMoralXplain, MFTCXplain, MOL, FactNews, and HausaHate. These resources support research on hate speech detection, moral reasoning, and automated fact-checking. Building on these resources, this thesis proposes five novel interpretable and explainable methods: Sentence-Level Factual Reasoning (SELFAR), Social Stereotype Analysis (SSA), Contextual Bag-of-Words with Human Rationales and Feature Optimization (B+M), Supervised Rational Attention (SRA), and Supervised Moral Rational Attention (SMRA). These approaches integrate human rationales, as well as moral and factual evidence into neural architectures, improving both predictive performance and explanation quality. Experiments across multiple datasets demonstrate consistent improvements over neural baselines while producing explanations that are more faithful and better aligned with human reasoning. Beyond its methodological contributions, the dissertation has generated significant scientific and societal impact. Its datasets and methods have been adopted by researchers from universities and research institutes in Brazil and abroad, including Microsoft Research and other leading institutions, supporting new studies on Explainable AI, Trustworthy NLP, and socially responsible language technologies for underrepresented languages. The quality and impact of this research have also been recognized through several national and international awards, including the Google LARA, the PROPOR Best Thesis Award, the Maria Carolina Monard Best Thesis Award in AI, the Trevisan Prize – AI for Good from Bocconi University, and the ACL D&I Award, as well as nominations for two Brazilian Computer Society (SBC) Thesis Awards in the Computer Science and Multimedia, Hypermedia, and Web categories.*

* **PhD Thesis:** <https://teses.usp.br/teses/disponiveis/55/55134/tde-07012025-155212/pt-br.html>
Google Scholar: <https://scholar.google.com/citations?user=42bzuFQAAAAJ&hl=en>
Lattes CV: <http://lattes.cnpq.br/0667506930148885>
Published Papers: <https://franciellevargas.github.io/>

Resumo. Esta tese de doutorado avança o estado da arte em Inteligência Artificial (IA) e Processamento de Linguagem Natural (PLN) ao abordar dois desafios críticos: a escassez de Corpora de alta qualidade para línguas sub-representadas no Sul Global e a interpretabilidade limitada dos modelos de linguagem modernos em tarefas com alto impacto social. A pesquisa concentra-se na detecção de discurso de ódio e na verificação automática de fatos, propondo novos conjuntos de dados, métodos explicáveis e benchmarks de avaliação para aprimorar a transparência, a equidade e a confiabilidade dos sistemas de PLN. A tese apresenta sete conjuntos de dados benchmarks para o português brasileiro e a língua hausa: HateBR, HateBRXplain, HateBRMoralXplain, MFTCXplain, MOL, FactNews e HausaHate. Esses recursos apoiam pesquisas em detecção de discurso de ódio, raciocínio moral das LLMs e verificação automática de fatos. Com base nesses recursos, a tese propõe cinco novos métodos interpretáveis e explicáveis: Sentence-Level Factual Reasoning (SELFAR), Social Stereotype Analysis (SSA), Contextual BoW with Human Rationales and Feature Optimization (B+M), Supervised Rational Attention (SRA) e Supervised Moral Rational Attention (SMRA). Essas abordagens integram justificativas humanas (human rationales), bem como evidências morais e factuais, às arquiteturas neurais, aprimorando tanto o desempenho preditivo quanto a qualidade das explicações. Experimentos realizados em múltiplos conjuntos de dados demonstram melhorias consistentes em relação a modelos neurais de referência, ao mesmo tempo em que produzem explicações mais fiéis e melhor alinhadas ao raciocínio humano. Além de suas contribuições metodológicas, a tese gerou impacto científico e social significativo. Seus conjuntos de dados e métodos foram adotados por pesquisadores de universidades e institutos de pesquisa no Brasil e no exterior, incluindo a Microsoft Research e outras instituições de referência, contribuindo para novos estudos em Inteligência Artificial Explicável (XAI), Processamento de Linguagem Natural Confiável (Trustworthy NLP) e tecnologias de linguagem socialmente responsáveis para línguas sub-representadas. A qualidade e o impacto desta pesquisa também foram reconhecidos por diversos prêmios nacionais e internacionais, incluindo o Google LARA, o Prêmio de Melhor Tese do PROPOR, o Prêmio Maria Carolina Monard de Melhor Tese em Inteligência Artificial, o Trevisan Prize – AI for Good da Universidade Bocconi na Itália, ACL D&I Awards, além de duas indicações a Prêmios de Teses da Sociedade Brasileira de Computação (SBC), nas categorias de Ciência da Computação e de Multimídia, Hipermídia e Web.

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 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, machine learning models may internalize and perpetuate normative social stereotypes, leading to systematic prediction errors (Davani et al. 2023). Another critical source of bias is the absence of contextual information (Kennedy et al. 2020). Similarly, given the prevalent practice in modern NLP of training AI systems on extensive corpora of human-generated text, representational biases can emerge. These biases manifest in various ways, such as an increased preference for European American names over African

*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.

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

American names (Caliskan et al. 2017), the association of more negative sentiments with individuals with disabilities (Hutchinson et al. 2020), and the reinforcement of ethnic stereotypes, such as linking Hispanic individuals with housekeepers or Asians with professors (Garg et al. 2018). 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 <u>Africans</u> , the most threatening kind of ethnic hatred is <u>black</u> against <u>black</u> . - New York Times	- hate speech
There is a great discrepancy between <u>whites</u> and <u>blacks</u> in SA. It is ... [because] <u>blacks</u> 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 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,

*<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

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.

2. Problem, Motivation and Objectives

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

3. Research Hypotheses

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

*<https://www.politifact.com/factchecks/2023/mar/29/william-barclay/syracuse-proposal-would-pay-gang-members-stay-away/>

4. Related Works

4.1. Artificial Intelligence for Automated Fact-Checking

Automated fact-checking relies on retrieving potentially relevant documents from various data sources (e.g., news websites, social media), determining the stance of each document regarding the input claim, and then predicting the claim’s factuality by aggregating the strength of these stances while considering the reliability of the source (Baly et al. 2018). In terms of methods, most existing systems classify the veracity of a claim, identify evidence, and predict whether the evidence supports or refutes the claim (Guo et al. 2022). Furthermore, automated fact-checking can be categorized into two main tasks: fact extraction and fact verification, both structured as a three-stage process: (i) identifying claims to check, (ii) finding appropriate evidence, and (iii) producing verdicts, as shown in Figure 2. Fact extraction, also known as claim detection, focuses on identifying the most relevant claims to check. Conversely, fact verification involves retrieving evidence from a corpus for a given claim whose veracity is unknown and predicting whether the evidence supports or refutes the claim (Guo et al. 2022). Typically, the input consists of a social media post or a document containing multiple claims. The output from claim detection is then used as input for claim verification (Guo et al. 2022). This process often focuses on sentence-level statements, a common practice among fact-checkers to ensure that only the context relevant to the claim is included (Mena 2019). Both tasks and methods are detailed below.

Claim detection relies on making a binary decision for each potential claim or ranking claims by importance (Nakov et al. 2018; Barrón-Cedeño et al. 2020). Additionally, claim detection depends on check-worthiness, meaning the identification of claims for which the general public would be interested in knowing the truth. For example, rumor detection considers factors such as language subjectivity and the growth of readership through social networks (Qazvinian et al. 2011). In both cases, the goal is to find factual and reliable content that is worth checking. For instance, (Barrón-Cedeño et al. 2020) created a set of auxiliary questions, such as “Does it contain a factual claim?” and “Is it of general interest?” to assist annotators in identifying check-worthy posts. Therefore, claim detection is typically framed as a classification task in which models predict whether claims are checkable or check-worthy (Guo et al. 2022). Earlier methods used supervised classifiers with feature engineering, relying on surface features such as Reddit karma and up-votes (Aker et al. 2017), Twitter-specific features (Enayet and El-Beltagy 2017), named entities and verbal forms in political transcripts (Zhong et al. 2020), or lexical and syntactic features (Zhou et al. 2020). Most recently, neural network approaches have been employed, including LSTM models (Kochkina et al. 2017; Tai et al. 2015), attention mechanisms (Guo et al. 2018), and graph neural networks (Yang et al. 2020).

Claim verification involves finding information beyond the claim itself (e.g., text, tables, knowledge bases, images, relevant metadata) that could potentially indicate veracity by determining whether the evidence supports or refutes the claim (Guo et al. 2022). A wide range of studies on claim verification have extracted veracity-indicative features, which reflect stances and writing styles from relevant texts such as news articles and microblog posts. These features have been used in traditional supervised models (Castillo et al. 2011; Ferreira and Vlachos 2016) and neural networks (Wang 2017; Ma et al. 2019; Wu et al. 2020) to learn the parameters. The most popu-

lar approaches are Recognizing Textual Entailment (RTE) and stance detection, which are used to predict whether evidence supports or refutes the claim (Guo et al. 2022). For example, stance detection is employed as a fine-grained filter to re-rank retrieved evidence (Thorne et al. 2018; Nie et al. 2019). In addition, retrieving multiple pieces of evidence and concatenating them into a single string, after which a textual entailment model is used, has also been employed to infer whether the evidence supports or refutes the claim (Schlichtkrull et al. 2021; Liu et al. 2020; Ma et al. 2019). Moreover, verifying whether a claim has already been fact-checked is also an essential step (Guo et al. 2022). Regarding this, (Hossain et al. 2020) verified if a tweet matched any known COVID-19 related misconceptions, classifying it as misinformation if a match was found. Finally, matching claims against previously verified ones can often be reduced to sentence-level similarity (Shaar et al. 2020), a method that has also been explored in the context of textual entailment (Guo et al. 2022).

4.2. Artificial Intelligence for Hate Speech Detection

hate speech detection systems aim to develop effective countermeasures to prevent the publication or spread of offensive or hateful content on the Web and social media platforms. These tasks are specifically concerned with identifying offensive language, toxicity, and hate speech targets (e.g., racism, misogyny, partyism). As noted by (Fortuna et al. 2020), expressions such as hate speech, offensive language, toxicity, and abusive language are often used interchangeably in the literature. Here, we consider hate speech detection to encompass two main tasks: offensive language and hate speech detection, as defined in detail as follows.

Offensive language detection is a binary classification task aimed at categorizing input data into offensive or non-offensive language. Offensive language includes any profanity, strongly impolite, rude, or vulgar expressions directed at individuals or groups, as defined by (Fortuna and Nunes 2018). This language can be explicit or implicit (Poletto et al. 2021), and encompasses insults, threats, and untargeted profanity (Zampieri et al. 2019), regardless of the target’s social identity (Vargas et al. 2024).

Hate speech detection is a binary classification task aimed at categorizing input data into hate speech (targeting groups based on social identity) or non-hate speech (non-target). Hate speech is a specific type of offensive language that can be explicit or implicit (Poletto et al. 2021; Vargas et al. 2021a). It targets and diminishes groups by inciting violence and hate based on characteristics as physical appearance, religion, ethnicity, or sexual orientation. It may appear in various linguistic styles, including subtle or humorous forms (Fortuna and Nunes 2018; Zampieri et al. 2019), which often uses stereotypes to express hate ideologies (Warner and Hirschberg 2012). In addition, hate speech tasks can focus on specific targets, such as *racism* (Ayele et al. 2023; Hasanuzzaman et al. 2017), *antisemitism* (Ozalp et al. 2020; Zannettou et al. 2020), *religious intolerance* (Ghosh Chowdhury et al. 2019), and *sexism* (Chiril et al. 2020; Guest et al. 2021; Almanea and Poesio 2022), or *aporophobia* (Kiritchenko et al. 2023). It is also categorized within toxicity detection tasks (Guimaraes et al. 2020; Xenos et al. 2021; Besspalov et al. 2023).

Considering the most relevant literature on hate speech detection, the main methods focus on the following approaches (Schmidt and Wiegand 2017): (i) *simple surface features*, using Bag-of-Words (BoW) models with unigram and n-grams

(Chen et al. 2012; Xu et al. 2012; Waseem and Hovy 2016); (ii) *word generalization*, using clustering methods such as LDA and word embeddings (Brown et al. 1992; Alharbi and Lee 2020); (iii) *sentiment analysis*, leveraging semantic polarity to predict offensive content (Van Hee et al. 2015; Njagi et al. 2015); (iv) *lexical resources*, using controlled vocabularies of offensive terms (Xiang et al. 2012; Vargas et al. 2021a); (v) *linguistic features*, incorporating Part-Of-Speech (POS) and syntactic structures (Zhong et al. 2016); (vi) *knowledge-based features*, using context and commonsense reasoning (Dinakar et al. 2012); (vii) *meta-information*, leveraging metadata to predict hate speech (Waseem and Hovy 2016); (viii) *multi-modal information*, which includes detecting hate speech in images, audio, and memes (Pramanick et al. 2021); (ix) *neural networks*, applying deep learning techniques such as CNNs and RNNs to extract features (Gambäck and Sikdar 2017); (x) *LLMs*, which fine-tune models or use prompt engineering to predict hate speech (Wulach et al. 2021; Plaza-del arco et al. 2023).

5. Research Methodology

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

6. Contributions and Results

This thesis introduces a comprehensive suite of benchmark resources, explainable methods, and computational tools that advance the state of the art in hate speech detection, automated fact-checking, and Explainable Artificial Intelligence (XAI), particularly for low-resource languages. It presents HateBR and HateBRXplain, the first large-scale expert-annotated datasets for hate speech detection in Brazilian Portuguese with human rationales for explainability; HausaHate, a benchmark for hate speech detection in Hausa, an underrepresented African language; MOL, the first multilingual offensive lexicon with contextual annotations and culturally adapted translations across six languages; and Fact-News, a benchmark dataset for sentence-level news source reliability estimation in Portuguese. The thesis also proposes several novel explainable NLP methods, including Sentence-Level Factual Reasoning (SELFAR) for explainable automated fact-checking, Social Stereotype Analysis (SSA) for counterfactual bias assessment through stereotypes and counter-stereotypes, and B+M, a contextualized bag-of-words model enhanced with interpretable feature optimization. These contributions are complemented by NoHateBrazil, a publicly available web-based system for explainable offensiveness analysis in Brazilian Portuguese. Collectively, these datasets, methods, and systems establish new benchmarks for future research in Portuguese and other underrepresented languages while advancing fundamental topics in Computer Science, including Natural Language Processing, Explainable AI, fairness-aware machine learning, and computational social science. The proposed methods improve model transparency, enable systematic bias detection and mitigation, and demonstrate that interpretability can be achieved without sacrificing predictive performance. By expanding NLP research beyond English-centric settings, this

thesis contributes to more inclusive and trustworthy AI technologies with practical applications in content moderation and automated fact-checking. Its scientific impact is evidenced by the adoption of its resources and methodologies by leading universities and research laboratories worldwide, including Microsoft Research, Carnegie Mellon University, and the University of Maryland, fostering new M.Sc. and Ph.D. research projects and opening opportunities for future technology transfer, including software registration and patentable innovations.

Due to the breadth of the results, not all findings and metrics could be included in this document. For the complete results and detailed metrics, please refer to the full dissertation: <https://teses.usp.br/teses/disponiveis/55/55134/tde-07012025-155212/pt-br.html>

7. Thesis Outcomes

The thesis generated several outcomes, including:

1. **15 (fifteen) published papers in top-tier international conferences:** In total, 15 (fifteen) papers were published in top-tier AI and NLP international conferences, workshops, and journals (e.g., EMNLP, NAACL, LREC, RANLP, NLP journal, AAAI, etc.) Papers: <https://franciellevargas.github.io/>
2. **Several benchmark datasets:** HateBR (Vargas et al. 2022), HateBRXplain (Salles et al. 2025), and HateBRMoralXplain (Vargas et al. 2026) (the first large-scale expert-annotated corpora for hate speech detection in Brazilian Portuguese, including hate speech and moral rationales for explainability); MOL (Vargas et al. 2025) (the first multilingual offensive lexicon, extracted from the HateBR containing 1,000 explicit and “clues” to identify implicit terms in Portuguese, translated into five languages while accounting for cultural aspects.); Fact-News (Vargas et al. 2023c) (a sentence-level annotated dataset for fact-checking in Portuguese); and MFTCXplain (Trager et al. 2025), the first benchmark for evaluating moral reasoning of LLMs.
3. **5 (Five) new post-hoc and self-explaining computational methods:** SEL-FAR (Vargas et al. 2023c) (the first explainable fact-checking method in Portuguese); SSA (Vargas et al. 2023a) (a counter-stereotype post-hoc explanation method to assess social bias in hate speech classifiers); B+M (Vargas et al. 2021b) (a contextual BoW with interpretable input and feature optimization for explainable hate speech detection); and SRA (Eilertsen et al. 2026) and SMRA (Vargas et al. 2026) (self-explaining methods that integrating hate speech and moral human-annotated rationales into deep learning models by attention alignment loss).
4. **A Web platform to combat hate speech in Brazil:** NoHateBrazil* (Vargas et al. 2023b), a public platform for hate speech detection in Brazilian Portuguese.
5. **Microsoft leveraged the HateBR dataset to train LLMs:** Microsoft Research has used the proposed HateBR dataset to train two 2 (two) LLMs: CultureLLM (Li et al. 2024a) and CulturePark (Li et al. 2024b), demonstrating its clear relevance and practical applicability.

*<http://143.107.183.175:14581/>

-
6. **Invited as an international visiting researcher:** The author of this thesis was invited to visit the University of Southern California (USC) and to present her work at an event of the Applied Social Media Lab at the *Berkman Klein Center for Internet & Society*, Harvard University.
 7. **Invited as an international keynote speaker:** The author of this thesis served as a keynote speaker at the Conference on Harmful Online Communication at the Leibniz Institute for the Social Sciences (GESIS), in Germany, and at the Conference cum Conclave on Emerging Trends in Journalistic and Media Practices at DG Vaishnav College, in India.
 8. **Invited roundtable lead at ICLR 2026 workshop:** The author of this thesis was invited to serve as a Roundtable Lead at the *Algorithmic Fairness Across Alignment Procedures and Agentic Systems (AFAA) Workshop*, held at the International Conference on Learning Representations 2026*.
 9. **Invited to top-tier NLP international conference and journals program committees:** The author of this thesis was invited to serve on the program committees of Language Resources and Evaluation, Expert Systems with Applications, Online Social Networks and Media, Natural Language Processing Journals, as well as EMNLP, ACL, NAACL, RANLP, COLING, LREC, ICWSM, WWW and CIKM conferences.
 10. **Co-organizer of top-tier international conferences and workshops:** The author of this thesis co-organized the prestigious ICWSM conference in 2021, 2022 and 2023, and was selected to organize ACL Workshop on Online Abuse and Harms (WOAH 2025) (Calabrese et al. 2025), and proposed the IJCNN 2025 Special Session on Explainable Deep Neural Networks for Responsible AI: Post-Hoc and Self-Explaining Approaches (DeepXplain 2025)[†].
 11. **More than 10 new Ph.D. and M.Sc. theses, and undergraduate research projects in Brazil:** Several prestigious public universities in Brazil have used our resources to propose new Ph.D. and M.Sc. theses, including institutions such as USP, UFMG, UFF, UFOP, UFC, among others.
 12. **Co-advisor of a computer science master’s student with a published paper in a top-tier NLP venue:** The author of this thesis co-advised a master’s student in computer science at DCC-UFMG, whose work was published at COLING 2025, a top-tier international NLP conference.
 13. **High citations from prestigious international institutions:** Research by prestigious institutions (e.g. Carnegie Mellon University, University of Maryland, Harvard University), has been significantly influenced by the resources provided in this thesis, as reflected in a substantial number of citations (**Google Scholar: 409 citations**)
 14. **Collaboration with researchers across 10 countries on 5 continents:** The author of this thesis has led projects and collaborated with researchers in ten different countries on five continents.
 15. **Press Coverage:** The results of this dissertation received relevant media coverage, with 10 reports, including international outlets, available in <https://francielleavargas.github.io/>.

*<https://www.afciworkshop.org/afaa-2026>

[†]<https://deepxplain.github.io/>

Acknowledgements

The authors gratefully acknowledge the AI4PG 2026 Grants Programme and GainForest for their financial support. This project was also supported by the Ministry of Science, Technology and Innovation through resources provided under Law No. 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. Additionally, this work was supported by the Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação (INCT-TILD-IAR), grant No. 408490/2024-1.

Referências

- [Aker et al. 2017] Aker, A., Derczynski, L., and Bontcheva, K. (2017). Simple open stance classification for rumour analysis. In Mitkov, R. and Angelova, G., editors, *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 31–39, Varna, Bulgaria.
- [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.
- [Alharbi and Lee 2020] Alharbi, A. I. and Lee, M. (2020). Combining character and word embeddings for the detection of offensive language in Arabic. In Al-Khalifa, H., Magdy, W., Darwish, K., Elsayed, T., and Mubarak, H., editors, *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 91–96, Marseille, France. European Language Resource Association.
- [Almanea and Poesio 2022] Almanea, D. and Poesio, M. (2022). ArMIS - the Arabic misogyny and sexism corpus with annotator subjective disagreements. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2282–2291, Marseille, France. European Language Resources Association.
- [Amazeen 2015] Amazeen, M. (2015). Revisiting the epistemology of fact-checking. *Critical Review*, 27(1):1–30.
- [Ayele et al. 2023] Ayele, A. A., Dinter, S., Yimam, S. M., and Biemann, C. (2023). Multilingual racial hate speech detection using transfer learning. In Mitkov, R. and Angelova, G., editors, *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 41–48, Varna, Bulgaria.
- [Baly et al. 2018] Baly, R., Mohtarami, M., Glass, J., Màrquez, L., Moschitti, A., and Nakov, P. (2018). Integrating stance detection and fact checking in a unified corpus. In Walker, M., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 21–27, New Orleans, Louisiana. Association for Computational Linguistics.
- [Barrón-Cedeño et al. 2020] Barrón-Cedeño, A., Elsayed, T., Nakov, P., Da San Martino, G., Hasanain, M., Suwaileh, R., Haouari, F., Babulkov, N., Hamdan, B., Nikolov, A., Shaar, S., and Ali, Z. S. (2020). Overview of checkthat! 2020: Automatic identification

-
- and verification of claims in social media. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings*, page 215–236, Berlin, Heidelberg.
- [Bespalov et al. 2023] Bespalov, D., Bhabesh, S., Xiang, Y., Zhou, L., and Qi, Y. (2023). Towards building a robust toxicity predictor. In Sitaram, S., Beigman Klebanov, B., and Williams, J. D., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 581–598, Toronto, Canada. Association for Computational Linguistics.
- [Brown et al. 1992] Brown, P. F., Della Pietra, V. J., deSouza, P. V., Lai, J. C., and Mercer, R. L. (1992). Class-based n -gram models of natural language. *Computational Linguistics*, 18(4):467–480.
- [Calabrese et al. 2025] Calabrese, A., de Kock, C., Nozza, D., Plaza-del Arco, F. M., Talat, Z., and Vargas, F., editors (2025). *Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH)*, Vienna, Austria. Association for Computational Linguistics.
- [Caliskan et al. 2017] Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.
- [Castillo et al. 2011] Castillo, C., Mendoza, M., and Poblete, B. (2011). Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web, WWW ’11*, page 675–684, New York, USA.
- [Chen et al. 2012] Chen, Y., Zhou, Y., Zhu, S., and Xu, H. (2012). Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing*, pages 71–80.
- [Chiril et al. 2020] Chiril, P., Moriceau, V., Benamara, F., Mari, A., Origgi, G., and Coulomb-Gully, M. (2020). An annotated corpus for sexism detection in French tweets. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1397–1403, Marseille, France. European Language Resources Association.
- [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.
- [Dinakar et al. 2012] Dinakar, K., Jones, B., Havasi, C., Lieberman, H., and Picard, R. (2012). Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems*, 2(3).
- [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, NY, USA. Association for Computing Machinery.
- [Eilertsen et al. 2026] Eilertsen, B., Björgfinsdóttir, R., Vargas, F., and Ramezani-Kebrya, A. (2026). Aligning attention with human rationales for self-explaining hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(44):37369–37378.
- [Enayet and El-Beltagy 2017] Enayet, O. and El-Beltagy, S. R. (2017). NileTMRG at SemEval-2017 task 8: Determining rumour and veracity support for rumours on Twitter. In Bethard, S., Carpuat, M., Apidianaki, M., Mohammad, S. M., Cer, D., and Jurgens, D., editors, *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 470–474, Vancouver, Canada. Association for Computational Linguistics.
- [Ferreira and Vlachos 2016] Ferreira, W. and Vlachos, A. (2016). Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page 1163–1168, San Diego, California.
- [Fortuna and Nunes 2018] Fortuna, P. and Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51(4):1–30.
- [Fortuna et al. 2020] Fortuna, P., Soler, J., and Wanner, L. (2020). Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6786–6794, Marseille, France. European Language Resources Association.
- [Gambäck and Sikdar 2017] Gambäck, B. and Sikdar, U. K. (2017). Using convolutional neural networks to classify hate-speech. In Waseem, Z., Chung, W. H. K., Hovy, D., and Tetreault, J., editors, *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90, Vancouver, BC, Canada. Association for Computational Linguistics.
- [Garg et al. 2018] Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.
- [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.
- [Ghosh Chowdhury et al. 2019] Ghosh Chowdhury, A., Didolkar, A., Sawhney, R., and Shah, R. R. (2019). ARHNet - leveraging community interaction for detection of religious hate speech in Arabic. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 273–280, Florence, Italy.
- [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.
- [Guest et al. 2021] Guest, E., Vidgen, B., Mittos, A., Sastry, N., Tyson, G., and Margetts, H. (2021). An expert annotated dataset for the detection of online misogyny. In

-
- Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1336–1350, Held Online.
- [Guimaraes et al. 2020] Guimaraes, S. S., Reis, J. C. S., Ribeiro, F. N., and Benevenuto, F. (2020). Characterizing toxicity on facebook comments in Brazil. In *Proceedings of the 26th Brazilian Symposium on Multimedia and the Web*, pages 253–260, Maranhão, Brazil.
- [Guo et al. 2018] Guo, H., Cao, J., Zhang, Y., Guo, J., and Li, J. (2018). Rumor detection with hierarchical social attention network. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM ’18, page 943–951, New York, NY, USA. Association for Computing Machinery.
- [Guo et al. 2022] Guo, Z., Schlichtkrull, M., and Vlachos, A. (2022). A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.
- [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.
- [Hasanuzzaman et al. 2017] Hasanuzzaman, M., Dias, G., and Way, A. (2017). Demographic word embeddings for racism detection on Twitter. In *Proceedings of the 8th International Joint Conference on Natural Language Processing*, pages 926–936, Taipei, Taiwan.
- [Hossain et al. 2020] Hossain, T., Logan IV, R. L., Ugarte, A., Matsubara, Y., Young, S., and Singh, S. (2020). COVIDLies: Detecting COVID-19 misinformation on social media. In Verspoor, K., Cohen, K. B., Conway, M., de Bruijn, B., Dredze, M., Mihalcea, R., and Wallace, B., editors, *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.
- [Hutchinson et al. 2020] Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., and Denuyl, S. (2020). Social biases in NLP models as barriers for persons with disabilities. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online. Association for Computational Linguistics.
- [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. Association for Computational Linguistics.
- [Kiritchenko et al. 2023] Kiritchenko, S., Curto Rex, G., Nejadgholi, I., and Fraser, K. C. (2023). Aporophobia: An overlooked type of toxic language targeting the poor. In Chung, Y.-l., Rottger, P., Nozza, D., Talat, Z., and Mostafazadeh Davani, A., editors, *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 113–125, Toronto, Canada. Association for Computational Linguistics.
- [Kochkina et al. 2017] Kochkina, E., Liakata, M., and Augenstein, I. (2017). Turing at SemEval-2017 task 8: Sequential approach to rumour stance classification with branch-LSTM. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 475–480, Vancouver, Canada. Association for Computational Linguistics.
- [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 (RDSM)*, pages 45–57, Barcelona, Spain (Online). Association for Computational Linguistics.
- [Li et al. 2024a] Li, C., Chen, M., Wang, J., Sitaram, S., and Xie, X. (2024a). Culturellm: Incorporating cultural differences into large language models. In *Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*.
- [Li et al. 2024b] Li, C., Teney, D., Yang, L., Wen, Q., Xie, X., and Wang, J. (2024b). Culturepark: Boosting cross-cultural understanding in large language models. In *Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*.
- [Liu et al. 2020] Liu, Z., Xiong, C., Sun, M., and Liu, Z. (2020). Fine-grained fact verification with kernel graph attention network. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7342–7351, Online. Association for Computational Linguistics.
- [Ma et al. 2019] Ma, J., Gao, W., Joty, S., and Wong, K.-F. (2019). Sentence-level evidence embedding for claim verification with hierarchical attention networks. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2561–2571, Florence, Italy. Association for Computational Linguistics.
- [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. Association for Computational Linguistics.
- [Mena 2019] Mena, P. (2019). Principles and boundaries of fact-checking: Journalists’ perceptions. *Journalism Practice*, 13(6):657–672.
- [Nakov et al. 2018] Nakov, P., Barrón-Cedeño, A., Elsayed, T., Suwaileh, R., Màrquez, L., Zaghoulani, W., Atanasova, P., Kyuchukov, S., and Da San Martino, G. (2018). Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims. In Bellot, P., Trabelsi, C., Mothe, J., Murtagh, F., Nie, J. Y., Soulier, L., SanJuan, E., Cappellato, L., and Ferro, N., editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 372–387, Cham. Springer International Publishing.
- [Nie et al. 2019] Nie, Y., Wang, S., and Bansal, M. (2019). Revealing the importance of semantic retrieval for machine reading at scale. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2553–2566, Hong Kong, China. Association for Computational Linguistics.
- [Nieminen and Rapeli 2019] Nieminen, S. and Rapeli, L. (2019). Fighting mispercepti-

-
- ons and doubting journalists’ objectivity: A review of fact-checking literature. *Political Studies Review*, 17(3):296–309.
- [Njagi et al. 2015] Njagi, D., Zuping, Z., Hanyurwimfura, D., and Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10:215–230.
- [Ozalp et al. 2020] Ozalp, S., Williams, M. L., Burnap, P., Liu, H., and Mostafa, M. (2020). Antisemitism on twitter: Collective efficacy and the role of community organisations in challenging online hate speech. *Social Media + Society*, 6(2):01–20.
- [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.
- [Plaza-del arco et al. 2023] Plaza-del arco, F. M., Nozza, D., and Hovy, D. (2023). Respectful or toxic? using zero-shot learning with language models to detect hate speech. In Chung, Y.-l., Rottger, P., Nozza, D., Talat, Z., and Mostafazadeh Davani, A., editors, *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 60–68, Toronto, Canada. Association for Computational Linguistics.
- [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.
- [Pramanick et al. 2021] Pramanick, S., Dimitrov, D., Mukherjee, R., Sharma, S., Akhtar, M. S., Nakov, P., and Chakraborty, T. (2021). Detecting harmful memes and their targets. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2783–2796, Online. Association for Computational Linguistics.
- [Qazvinian et al. 2011] Qazvinian, V., Rosengren, E., Radev, D. R., and Mei, Q. (2011). Rumor has it: Identifying misinformation in microblogs. In Barzilay, R. and Johnson, M., editors, *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1589–1599, Edinburgh, Scotland, UK. Association for Computational Linguistics.
- [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.
- [Schlichtkrull et al. 2021] Schlichtkrull, M. S., Karpukhin, V., Oguz, B., Lewis, M., Yih, W.-t., and Riedel, S. (2021). Joint verification and reranking for open fact checking

-
- over tables. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6787–6799, Held Online. Association for Computational Linguistics.
- [Schmidt and Wiegand 2017] Schmidt, A. and Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the 5th International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain.
- [Shaar et al. 2020] Shaar, S., Babulkov, N., Da San Martino, G., and Nakov, P. (2020). That is a known lie: Detecting previously fact-checked claims. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3607–3618, Held Online. Association for Computational Linguistics.
- [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.
- [Tai et al. 2015] Tai, K. S., Socher, R., and Manning, C. D. (2015). Improved semantic representations from tree-structured long short-term memory networks. In Zong, C. and Strube, M., editors, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566, Beijing, China. Association for Computational Linguistics.
- [Tajfel 1979] Tajfel, H. (1979). An integrative theory of intergroup conflict. *The social psychology of intergroup relations/Brooks/Cole*.
- [Thorne et al. 2018] Thorne, J., Vlachos, A., Cocarascu, O., Christodoulopoulos, C., and Mittal, A. (2018). The fact extraction and VERification (FEVER) shared task. In Thorne, J., Vlachos, A., Cocarascu, O., Christodoulopoulos, C., and Mittal, A., editors, *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.
- [Trager et al. 2025] Trager, J., Vargas, F., Alves, D., Guida, M., Ngueajio, M. K., Agrawal, A., Daryani, Y., Malekabi, F. K., and Plaza-del Arco, F. M. (2025). MFTCXplain: A multilingual benchmark dataset for evaluating the moral reasoning of LLMs through multi-hop hate speech explanation. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V., editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15709–15740, Suzhou, China. Association for Computational Linguistics.
- [Tsveltkov et al. 2019] Tsveltkov, 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, NY, USA. Association for Computing Machinery.
- [Van Hee et al. 2015] Van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., De Pauw, G., Daelemans, W., and Hoste, V. (2015). Detection and fine-grained classification of cyberbullying events. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 672–680, Hissar, Bulgaria.

-
- [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. 2024] Vargas, F., Carvalho, I., Pardo, T., and Benevenuto, F. (2024). Context-aware and expert data resources for brazilian portuguese hate speech detection. *Natural Language Processing*, pages 1–22.
- [Vargas et al. 2025] Vargas, F., Carvalho, I., Pardo, T. A. S., and Benevenuto, F. (2025). Context-aware and expert data resources for brazilian portuguese hate speech detection. *Natural Language Processing*, 31(2):435–456.
- [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. 2021a] Vargas, F., Goes, G., Carvalho, I., Benevenuto, F., and Pardo, T. (2021a). Contextual-lexicon approach for abusive language detection. In *Proceedings of the Recent Advances in Natural Language Processing*, pages 1438–1447, Held Online.
- [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. 2021b] Vargas, F., Rodrigues de Góes, F., Carvalho, I., Benevenuto, F., and Pardo, T. (2021b). Contextual-lexicon approach for abusive language detection. In Mitkov, R. and Angelova, G., editors, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1438–1447, Held Online. INCOMA Ltd.
- [Vargas et al. 2026] Vargas, F., Trager, J., Alves, D., Guida, M., Thapa, S., Atıl, B., Dementieva, D., Smart, A. J., and Agrawal, A. (2026). Self-explaining hate speech detection with moral rationales. In Liakata, M., Moreira, V. P., Zhang, J., and Jurgens, D., editors, *Findings of the Association for Computational Linguistics: ACL 2026*, pages 34109–34131, San Diego, California, United States. Association for Computational Linguistics.
- [Wang 2017] Wang, W. Y. (2017). “liar, liar pants on fire”: A new benchmark dataset for fake news detection. In Barzilay, R. and Kan, M.-Y., editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 422–426, Vancouver, Canada.
- [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.

-
- [Warner and Hirschberg 2012] Warner, W. and Hirschberg, J. (2012). Detecting hate speech on the world wide web. In *Proceedings of the 2nd Workshop on Language in Social Media*, pages 19–26, Montréal, Canada.
- [Waseem and Hovy 2016] Waseem, Z. and Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics - Student Research Workshop*, pages 88–93, California, USA.
- [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.
- [Wu et al. 2020] Wu, L., Rao, Y., Zhao, Y., Liang, H., and Nazir, A. (2020). DTCA: Decision tree-based co-attention networks for explainable claim verification. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1024–1035, Held Online. Association for Computational Linguistics.
- [Wullach et al. 2021] Wullach, T., Adler, A., and Minkov, E. (2021). Fight fire with fire: Fine-tuning hate detectors using large samples of generated hate speech. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4699–4705, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- [Xenos et al. 2021] Xenos, A., Pavlopoulos, J., and Androutsopoulos, I. (2021). Context sensitivity estimation in toxicity detection. In Mostafazadeh Davani, A., Kiela, D., Lambert, M., Vidgen, B., Prabhakaran, V., and Waseem, Z., editors, *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 140–145, Online. Association for Computational Linguistics.
- [Xiang et al. 2012] Xiang, G., Fan, B., Wang, L., Hong, J., and Rose, C. (2012). Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, page 1980–1984, New York, United States.
- [Xu et al. 2012] Xu, J.-M., Jun, K.-S., Zhu, X., and Bellmore, A. (2012). Learning from bullying traces in social media. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 656–666, Montréal, Canada.
- [Yang et al. 2020] Yang, X., Lyu, Y., Tian, T., Liu, Y., Liu, Y., and Zhang, X. (2020). Rumor detection on social media with graph structured adversarial learning. In Bessiere, C., editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 1417–1423. International Joint Conferences on Artificial Intelligence Organization. Main track.
- [Zampieri et al. 2019] Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019). Predicting the type and target of offensive posts in social media. In *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1415–1420, Minnesota, USA.

-
- [Zannettou et al. 2020] Zannettou, S., Finkelstein, J., Bradlyn, B., and Blackburn, J. (2020). A quantitative approach to understanding online antisemitism. In *Proceedings of the 14th International AAAI Conference on Web and Social Media*, pages 786–797, Online.
- [Zhong et al. 2016] Zhong, H., Li, H., Squicciarini, A., Rajtmajer, S., Griffin, C., Miller, D., and Caragea, C. (2016). Content-driven detection of cyberbullying on the instagram social network. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, page 3952–3958, New York, United States.
- [Zhong et al. 2020] Zhong, W., Xu, J., Tang, D., Xu, Z., Duan, N., Zhou, M., Wang, J., and Yin, J. (2020). Reasoning over semantic-level graph for fact checking. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6170–6180, Held Online. Association for Computational Linguistics.
- [Zhou et al. 2020] Zhou, X., Jain, A., Phoha, V. V., and Zafarani, R. (2020). Fake news early detection: A theory-driven model. *Digital Threats*, 1(2).