Context Matters, Bias Matters Too: Assessing Fairness in Profile Embeddings in the HateBR Dataset

Vinícius Moitinho

[1]Universidade Federal de Sergipe, Sergipe, Brasil
vmoitinhoss@gmail.com

Abstract. *Offensive language detection is challenged by discourse's contextual nature, especially in Brazilian political contexts. This work examines how domain-specific information affects classifying offensiveness toward public figures, using the HateBR corpus. We compare three BERTimbau-based approaches: a domain-agnostic baseline, a model with textual metadata injection, and a hybrid with latent profile embeddings. Results show profile embeddings outperform other approaches in accuracy and convergence, indicating the model learns target-specific patterns. However, fairness analysis reveals a trade-off: specialization increases False Positive Rate variance across profiles, suggesting the model absorbs dataset biases. Findings reinforce the need for models integrating equity and context sensitivity in online debate.*

1. Introduction

The widespread adoption of social media has profoundly transformed the dynamics of public debate, establishing direct and continuous channels of interaction between citizens and political representatives. While this transformation has expanded opportunities for political participation and democratic engagement, it has also facilitated the large-scale dissemination of hostile and abusive discourse. As a result, the automatic detection and moderation of offensive content has become a critical challenge for preserving healthy democratic deliberation in online environments [Davidson et al. 2017].

Recent advances in Natural Language Processing (NLP), particularly the emergence of Transformer-based architectures such as BERT [Devlin et al. 2018], have substantially improved performance in text classification tasks, including hate speech and offensive language detection. These models leverage contextualized representations that capture complex linguistic patterns and long-range dependencies, enabling significant gains over traditional machine learning approaches [Fortuna and Nunes 2018]. Nevertheless, the detection of offensiveness in political discourse remains a particularly challenging problem. Political language is intrinsically contextual, often involving irony, sarcasm, implicit references, and pragmatic cues that are difficult to capture through purely lexical or syntactic signals.

Beyond the inherent linguistic complexity of the task, another major challenge concerns the issue of algorithmic fairness. Deep learning models trained on large corpora tend to reproduce and sometimes amplify statistical regularities present in the training data, which may lead to systematic disparities in model behavior across different social groups [Sap et al. 2019]. In the context of hate speech detection, such biases may manifest when models incorrectly associate certain identities, dialects, or demographic attributes with toxicity or offensive content.

These concerns become especially salient in political contexts. Discursive aggression in online political debate is rarely uniformly distributed; instead, it is often shaped by ideological polarization and the identity of the targeted actor. In such scenarios, models that incorporate information about the political figure mentioned in a comment may inadvertently learn statistical shortcuts that associate specific individuals with offensive labels. This phenomenon can lead to what we refer to as *prevalence bias*, where the classifier relies on the historical frequency of hostility toward certain politicians rather than performing a proper semantic analysis of the textual content.

At the same time, contextual information may also provide valuable signals that improve classification performance. Recent research suggests that incorporating metadata or contextual representations can enhance hate speech detection systems, particularly in domains where meaning strongly depends on situational context. This raises an important methodological dilemma: how can models leverage contextual knowledge without introducing unfair biases tied to the identity of the target?

This work investigates precisely this trade-off between contextual awareness and predictive fairness in automated moderation systems. Our study begins with a statistical examination of the HateBR corpus, where we apply a rigorous **binomial hypothesis test** to assess whether the distribution of offensive comments is statistically dependent on the political profile of the target. The results confirm a significant association between these variables, indicating that contextual identity information may influence model learning dynamics.

Based on this premise, we design and compare three architectures built upon the BERTimbau model:

1. A unimodal, domain-agnostic baseline that relies solely on textual information;
2. An explicit contextual configuration that incorporates metadata through textual concatenation;
3. A hybrid architecture that integrates trainable **latent profile embeddings**, allowing the model to represent contextual attributes in a structured latent space.

The central objective of this study is to analyze how different representations of the target entity influence both predictive performance and fairness. In particular, we evaluate variations in overall accuracy as well as disparities in False Positive Rates (FPR) and False Negative Rates (FNR) across the six political profiles present in the dataset. By doing so, we aim to provide empirical evidence regarding the feasibility of moderation systems that are simultaneously context-aware and socially responsible.

2. Related Work

The automatic detection of hate speech and offensive language on social media has become a central topic in Natural Language Processing research, largely due to the rapid growth of user-generated content and the societal impacts associated with the dissemination of discriminatory messages. Early approaches relied primarily on lexicon-based techniques and rule-based filtering systems, which attempted to identify offensive content through predefined lists of profane or abusive expressions. However, several studies demonstrated the limitations of these approaches, particularly their inability to capture contextual nuances, figurative language, or implicit forms of aggression

[Chen et al. 2012, Schmidt and Wiegand 2017]. As a result, such methods frequently exhibited high false positive rates and limited generalization capabilities.

To overcome these limitations, subsequent research explored the use of supervised machine learning models incorporating lexical, syntactic, and stylistic features. These approaches employed classifiers such as Support Vector Machines and Logistic Regression combined with engineered linguistic features, including part-of-speech patterns and dependency structures. Such feature-based models provided moderate improvements in performance but still struggled with semantic ambiguity and domain transferability.

The emergence of distributed word representations marked a significant advancement in the field. Word embedding techniques such as Word2Vec, GloVe, and FastText enabled models to capture semantic similarity between words and reduce sparsity issues inherent in traditional bag-of-words representations. Building upon these representations, several studies adopted deep neural architectures for hate speech detection. For example, [Badjatiya et al. 2017] applied Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs) to classify offensive tweets, reporting substantial improvements over traditional machine learning baselines, with gains exceeding 18 points in F1-score.

More recently, the introduction of contextualized language models has further advanced the state of the art in offensive language detection. Transformer-based architectures such as BERT [Devlin et al. 2018] generate dynamic word representations conditioned on their surrounding context, allowing models to better handle informal language, sarcasm, and complex discourse structures commonly found in social media. Several studies have successfully applied these models to hate speech detection tasks. For instance, [Dai et al. 2020] proposed a multitask learning framework based on BERT that achieved competitive results in the SemEval-2020 shared task on offensive language identification.

In the Brazilian context, parallel research efforts have focused on addressing the scarcity of annotated corpora and language resources tailored to Portuguese. [Pelle and Moreira 2017] introduced the OFFCOMBR dataset and evaluated traditional supervised classifiers for offensive language detection in Portuguese, highlighting the importance of culturally and linguistically adapted resources. More recently, datasets such as HateBR have further contributed to advancing research in this area by providing large-scale annotated corpora for studying abusive language in Brazilian social media.

This line of research has continued to evolve, with several recent works further exploring the HateBR corpus and related phenomena in Portuguese-language discourse. [Silva et al. 2026] proposed a multitask Transformer architecture for joint offensive language detection and target identification on HateBR, demonstrating that jointly modeling these tasks improves performance over single-task baselines, although their approach does not explicitly model latent representations of the targeted profiles nor assess fairness across them. [Parafino et al. 2025] conducted a preliminary study on the co-occurrence between hate speech and fake news in Portuguese texts, reinforcing the relevance of contextual and pragmatic factors—such as the surrounding informational environment—in the manifestation of offensive discourse. Complementarily, [de Araújo Miranda and de Oliveira Rodrigues 2025] proposed an integrated approach for hate speech detection in social media combining textual vectorization with emoji

representations, illustrating the growing interest in incorporating multimodal or extralinguistic signals beyond raw text to improve classification robustness.

Another promising direction explored in recent literature involves hybrid representation approaches that combine multiple embedding sources or incorporate contextual metadata. For example, [Badri et al. 2022] integrated GloVe and FastText embeddings into a BiGRU architecture, achieving approximately 90% F1-score on the OLID dataset. These approaches suggest that combining complementary sources of information can improve the robustness of offensive language classifiers.

Despite these advances, several challenges remain unresolved. As emphasized by [Talat and Hovy 2016], the perception and labeling of offensive language are inherently subjective and strongly influenced by social, cultural, and contextual factors. Consequently, models trained on existing datasets may exhibit biases or fail to generalize across domains, particularly in politically polarized environments.

Taken together, these studies illustrate the methodological evolution of hate speech detection, moving from rule-based systems to deep contextual and multitask architectures. However, even recent works on HateBR [Silva et al. 2026] focus primarily on accuracy and joint task performance, leaving open questions regarding how the identity of the targeted public figure shapes discursive patterns and, more critically, how fairness metrics behave when models specialize on a per-target basis. This study addresses this gap by explicitly modeling latent profile embeddings and conducting a fairness-oriented analysis of False Positive Rate variance across targets, an aspect that remains underexplored in the current Brazilian Portuguese literature.

3. Dataset

The dataset used in this study is **HateBR**, a publicly available corpus designed for the detection of offensive language and hate speech in Brazilian Portuguese [Haddad et al. 2021]. The dataset contains approximately 7,000 comments collected from Instagram posts associated with Brazilian political figures. All instances were manually annotated by trained annotators, ensuring a high level of reliability and methodological consistency. HateBR represents an important contribution to the field, addressing the scarcity of large-scale annotated resources tailored to the sociolinguistic characteristics of Brazilian Portuguese.

The annotation scheme of HateBR follows a hierarchical structure that enables different levels of analysis:

- **Binary Level:** Classification of comments as offensive or non-offensive;
- **Offensiveness Intensity:** Categorization of offensive messages according to their severity (low, moderate, and high);
- **Hate Category:** Identification of specific hate targets, including racism, sexism, homophobia, and political hostility.

In this work, we focus exclusively on the **binary classification level**, which distinguishes offensive from non-offensive comments. This simplification allows us to concentrate on the relationship between textual content and the political identity of the target.

The quantitative distribution of messages across the monitored political profiles is presented in Table 1. In the context of this study, each political profile is treated as

Tabela 1. Distribution of offensive and non-offensive labels by Instagram profile.

Profile	Non-Offensive	Offensive	Total
Carla Zambelli	491	990	1481
Eduardo Bolsonaro	737	414	1151
Fernando Haddad	854	264	1118
Gleisi Hoffmann	395	439	834
Joice Hasselmann	463	695	1158
Manuela d'Ávila	560	698	1258
Total	**3500**	**3500**	**7000**

an **analytical domain**, enabling the investigation of how offensive discourse manifests across different spheres of political debate.

As illustrated in Figure 1, there is considerable variation in both the volume and proportion of offensive messages across profiles. While some domains present a relatively balanced distribution between classes, others exhibit a strong predominance of offensive content. This heterogeneity suggests that the prevalence of toxicity is not uniformly distributed across political actors.

Such inter-domain variability reinforces the importance of context-aware modeling approaches. In particular, it motivates the exploration of architectures capable of learning **target-specific representations**, allowing the model to capture contextual patterns associated with each political profile. This idea underlies the modeling strategies proposed in the Methodology section.

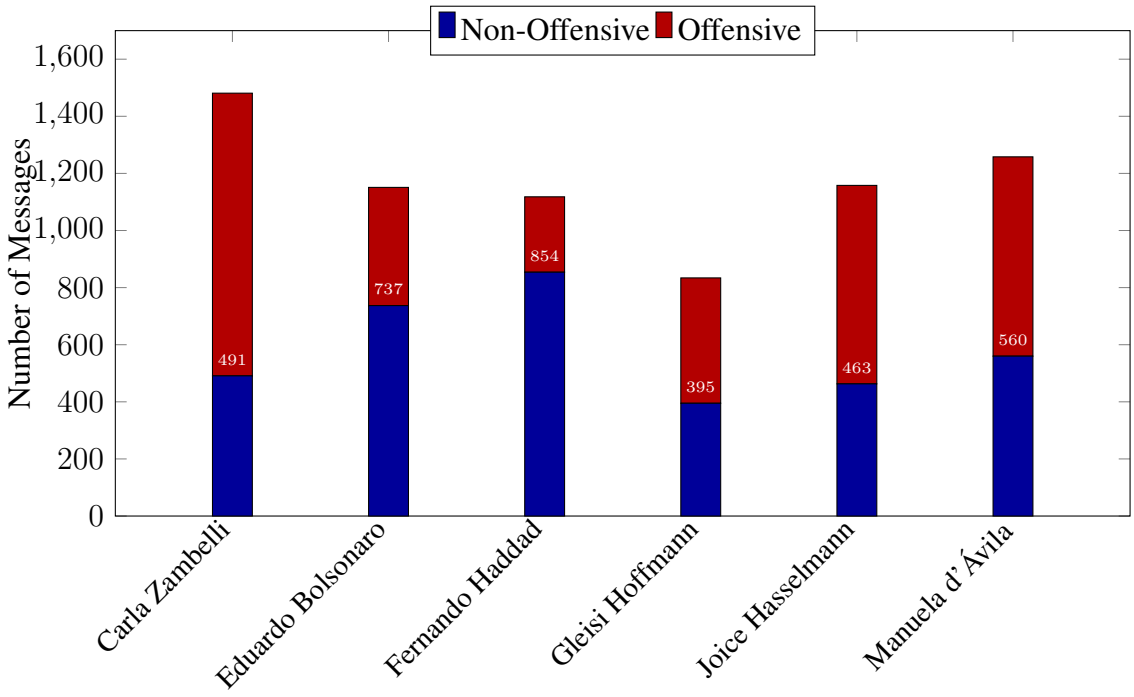

Figura 1. Distribution of offensive and non-offensive messages by political profile.

4. Statistical Analysis of Offensiveness Distribution

To investigate whether the incidence of offensive language depends on the targeted political profile, we conducted a binomial hypothesis test for each domain present in the dataset. Under the null hypothesis (H_0), the occurrence of offensive messages is assumed to be independent of the political target. Given that the HateBR dataset is globally balanced, with equal numbers of offensive and non-offensive messages, the expected probability of an offensive comment under H_0 is $p = 0.5$.

Thus, for each domain containing n messages, the number of offensive instances is modeled as a binomial random variable $X \sim Binomial(n, 0.5)$. Deviations from this expected proportion may indicate that the prevalence of offensive content is conditioned by the identity of the targeted profile rather than arising from random variation.

The statistical evaluation adopts a significance level of $\alpha = 0.05$. P-values below this threshold provide evidence to reject the null hypothesis, suggesting that the observed proportion of offensive comments significantly differs from the expected value.

Table 2 summarizes the cumulative probabilities obtained for each domain. The value $p_{\leq}$ represents the probability of observing a number of offensive messages equal to or lower than the recorded value under the null hypothesis, while $p_{\geq}$ denotes the probability of observing a number equal to or greater than the observed count. Consequently, small values of $p_{\leq}$ indicate a statistically significant *under-representation* of offensive messages, whereas small values of $p_{\geq}$ indicate a statistically significant *over-representation*.

Tabela 2. Results of the binomial test by domain.

Domain	$p_{\leq}$	$p_{\geq}$
Carla Zambelli	0.99	2.54×10^{-39}
Eduardo Bolsonaro	6.23×10^{-22}	0.99
Fernando Haddad	2.79×10^{-73}	1.00
Gleisi Hoffmann	0.94	0.06
Joice Hasselmann	0.99	4.86×10^{-12}
Manuela d'Ávila	0.99	5.53×10^{-5}

The results reveal substantial disparities across political domains. The profiles of **Carla Zambelli**, **Joice Hasselmann**, and **Manuela d'Ávila** exhibit extremely small right-tailed probabilities ($p_{\geq}$), indicating a statistically significant over-representation of offensive messages relative to the expected baseline. In contrast, the domains associated with **Eduardo Bolsonaro** and **Fernando Haddad** present extremely small left-tailed probabilities ($p_{\leq}$), suggesting a significantly lower prevalence of offensive comments than expected.

The profile of **Gleisi Hoffmann**, however, does not exhibit statistically significant deviations from the expected distribution, indicating that the proportion of offensive messages is consistent with the overall dataset balance.

Taken together, these findings demonstrate that the distribution of offensive language varies significantly across political profiles. This statistical heterogeneity suggests

that the identity of the targeted actor may influence the prevalence of hostile discourse. Such variation motivates the investigation of context-aware modeling strategies and raises important considerations regarding potential bias and fairness in automated moderation systems.

5. Experimental Methodology

This section describes the experimental configurations designed to evaluate how contextual information related to the target identity affects offensive language detection.

5.1. Experimental Protocol

To ensure statistical robustness and reproducibility, all experiments followed a consistent training and evaluation protocol. The dataset was partitioned using a stratified random split, allocating 70% of the data for training, 15% for validation, and 15% for testing.

Given the sensitivity of Transformer-based models to weight initialization and training stochasticity, each experimental configuration was executed across five independent runs using different random seeds. The results reported in the subsequent sections correspond to the mean and standard deviation of the evaluation metrics obtained across these runs.

All models employ *bert-base-portuguese-cased* (BERTimbau) as the base encoder. The model was fine-tuned end-to-end on the HateBR dataset, allowing the upper layers of the Transformer to adapt to the linguistic and pragmatic characteristics of Brazilian political discourse.

5.2. Experiment 1: Domain-Agnostic Text Classification Baseline

The first experiment establishes a baseline model based solely on textual information, following the standard Transformer-based sequence classification paradigm. The objective is to evaluate the intrinsic capacity of BERTimbau to detect offensive language without incorporating contextual information about the political target.

In this configuration, the input consists exclusively of the tokenized message. The sentence representation is obtained from the final hidden state corresponding to the special `[CLS]` token, denoted as $\mathbf{h}_{CLS} \in R^d$.

The probability of offensiveness $\hat{y}$ is obtained through a linear classification layer followed by a sigmoid activation:

$$\hat{y} = \sigma(\mathbf{W}\mathbf{h}_{CLS} + \mathbf{b}) \tag{1}$$

where $\mathbf{W} \in R^{k \times d}$ represents the weight matrix of the classification layer, k denotes the number of classes, and $\mathbf{b}$ is the bias vector.

Because this model does not incorporate contextual signals related to the target identity, its predictions rely entirely on lexical and semantic patterns present in the text. Consequently, the model may struggle with instances where offensiveness depends on pragmatic or domain-specific political references.

5.3. Experiment 2: Contextualization via Metadata Injection

The second experiment introduces contextual information by explicitly incorporating the identity of the political profile into the input sequence. The goal is to enable the Transformer's self-attention mechanism to capture potential interactions between the message content and the identity of the target.

In this configuration, the input sequence is modified as follows:

```
[CLS] Profile_ID [SEP] Message [SEP]
```

Under this formulation, the contextual representation becomes conditional on the target identity and can be expressed as

$$\mathbf{h}'_{CLS} = f_\theta(Text \mid Identity) \tag{2}$$

This strategy may improve the model's ability to detect target-specific attack patterns. However, it also introduces potential risks related to **prevalence bias**. In particular, the classifier may learn statistical shortcuts by associating specific identities with higher probabilities of offensiveness due to imbalances in the training data, rather than relying on the semantic content of the message.

5.4. Experiment 3: Latent Profile Embeddings

The third experiment proposes a hybrid architecture in which the identity of the political profile is represented as a trainable dense embedding. Unlike the previous approach, this method separates the textual representation from the identity representation while allowing both components to interact at the decision layer.

The architecture contains two feature extraction components:

1. The Transformer encoder generates the textual representation $\mathbf{h}_{CLS} \in R^d$.
2. A categorical embedding layer maps each profile identifier to a trainable vector $\mathbf{e}_{profile} \in R^m$.

The final classification layer operates on the concatenation of these representations:

$$\hat{y} = \sigma\left(\mathbf{W}_{out}[\mathbf{h}_{CLS} \oplus \mathbf{e}_{profile}] + \mathbf{b}\right) \tag{3}$$

where $\oplus$ denotes vector concatenation.

During training, the profile embeddings are jointly optimized with the parameters of the language model through backpropagation. This configuration enables the model to learn latent representations that capture similarities between political profiles in terms of the discourse patterns associated with them.

Nevertheless, the explicit inclusion of identity representations raises important fairness considerations. If the learned embeddings encode imbalances present in the training data, the model may exhibit systematic disparities across profiles, such as higher False Positive Rates (FPR) for specific political figures. For this reason, the evaluation protocol explicitly analyzes performance variations across domains.

6. Results and Analysis

The models were evaluated using three metrics: Accuracy, False Positive Rate (FPR), and False Negative Rate (FNR). Table 3 reports the mean and standard deviation obtained across five independent runs with different random seeds for each experimental configuration.

Tabela 3. Performance comparison among the proposed architectures (Mean $\pm$ Standard Deviation).

Experiment	Accuracy	FPR	FNR
Experiment 1	0.9089 ± 0.0071	$\mathbf{0.0732 \pm 0.0154}$	0.1544 ± 0.0277
Experiment 2	0.9142 ± 0.0069	0.0908 ± 0.0223	$\mathbf{0.1511 \pm 0.0262}$
Experiment 3	$\mathbf{0.9167 \pm 0.0052}$	0.0783 ± 0.0132	0.1753 ± 0.0440

6.1. Performance and Fairness Considerations

The experimental results indicate that incorporating contextual information improves overall predictive performance. In particular, the architecture based on **Latent Profile Embeddings** (Experiment 3) achieved the highest mean accuracy (0.9167) and the lowest variance across random seeds ($\sigma = 0.0052$), suggesting greater training stability.

This result supports the hypothesis that modeling the political target as a latent variable provides additional contextual signals that help the encoder disambiguate offensive language in politically nuanced discourse.

However, when analyzing the error metrics, important trade-offs emerge.

- **Latent Embedding Model (Exp. 3):** Although this architecture achieves the highest overall accuracy, it also presents the highest False Negative Rate (0.1753) and the largest variance for this metric. This suggests that the model may become more prone to overlooking certain offensive messages, particularly in domains where contextual cues are more ambiguous.
- **Metadata Injection (Exp. 2):** The explicit concatenation of the target identity produces the highest False Positive Rate (0.0908). This behavior is consistent with the hypothesis of *prevalence bias*, where the model may associate specific identities with higher probabilities of offensiveness due to patterns observed during training.
- **Text-Only Baseline (Exp. 1):** The domain-agnostic baseline achieves the lowest False Positive Rate (0.0732). This indicates that, in the absence of explicit contextual signals, the model relies more strictly on textual semantics when identifying offensive content.

Taken together, these results illustrate the trade-off between contextual specialization and error distribution. While contextual modeling improves overall accuracy, it may also introduce new sources of bias related to the identity of the target.

From a fairness perspective, these findings motivate the analysis of error disparities across political domains. In particular, variations in False Positive and False Negative Rates across profiles may indicate that the model provides uneven levels of protection against offensive language. The following section further investigates this issue by examining error distributions across the six political profiles present in the dataset.

7. Conclusion

This study examined the impact of incorporating contextual information into the detection of offensive language directed at Brazilian political figures. Starting from a statistical analysis of the HateBR dataset, we first demonstrated that the distribution of offensive messages varies significantly across political profiles. This finding motivated the investigation of context-aware modeling strategies capable of capturing domain-specific discourse patterns.

To this end, we compared three Transformer-based architectures that differ in how contextual information about the political target is incorporated. The experimental results indicate that modeling the target identity through **latent profile embeddings** (Experiment 3) yields the best overall performance, achieving the highest accuracy (91.67%) and the lowest variance across random seeds. These results suggest that representing the political target as a latent variable provides useful contextual signals that help the model disambiguate offensive language in politically nuanced discourse.

However, the analysis of error metrics revealed an important trade-off between contextual specialization and fairness. In particular, the embedding-based architecture exhibited a higher False Negative Rate (FNR) and greater variance for this metric, suggesting that the model may overlook certain offensive messages in specific domains. In addition, the metadata-injection strategy (Experiment 2) produced the highest False Positive Rate (FPR), providing empirical evidence of *prevalence bias*, where the explicit presence of a political identity may lead the model to overestimate the probability of offensiveness.

These findings highlight the importance of carefully designing context-aware moderation systems, as the inclusion of identity-related information can simultaneously improve predictive performance while introducing potential biases across targets.

As future work, we plan to investigate fairness-aware training strategies aimed at reducing disparities in error rates across political profiles. In particular, incorporating regularization mechanisms or fairness-aware loss functions may help mitigate the effects of target-based bias during the learning process. Furthermore, extending the dataset to include a broader diversity of political actors and ideological contexts would allow for a more comprehensive evaluation of how contextual representations influence offensive language detection in polarized environments.

Overall, this work contributes to the understanding of how contextual information interacts with predictive performance and fairness in automated moderation systems for political discourse.

Referências

Badjatiya, P., Gupta, S., Gupta, M., and Varma, V. (2017). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760.

Badri, N., Kboubi, F., and Chaibi, A. H. (2022). Combining fasttext and glove word embeddings for offensive and hate speech text detection. *Procedia Computer Science*, 207:769–778.

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*, pages 71–80. IEEE.

Dai, W., Li, W., Li, R., Ji, D., and Li, C. (2020). Kungfupanda at semeval-2020 task 12: Bert-based multi-task learning for offensive language detection. *arXiv preprint arXiv:2004.13432*.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM)*, pages 512–515.

de Araújo Miranda, A. L. and de Oliveira Rodrigues, C. M. (2025). Uma abordagem integrada para detecção de discurso de ódio em mídias sociais utilizando vetorização de textos e emojis. In *Anais do Workshop sobre as Implicações da Computação na Sociedade (WICS)*, pages 247–255. SBC.

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

Fortuna, P. and Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51(4):1–30.

Haddad, D., Monteiro, R., and Gonçalves, T. (2021). Hatebr: An annotated corpus for hate speech detection in brazilian portuguese. In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology (STIL)*.

Parafino, C., Tomaszewski, J., de Freitas, L. A., and Santana, B. S. (2025). Discursos de ódio e notícias falsas: Estudo preliminar da coocorrência entre esses fenômenos em textos em português. In *Anais da Escola Regional de Aprendizado de Máquina e Inteligência Artificial da Região Sul (ERAMIA-RS)*, pages 180–183. SBC.

Pelle, R. M. and Moreira, V. P. (2017). Offensive comments in the brazilian web: a dataset and analysis. In *Proceedings of the 8th Brazilian Symposium on Information and Human Language Technology (STIL)*.

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.

Schmidt, A. and Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10.

Silva, G., Silva, P., Peixoto, M., Moreira, G., and Luz, E. (2026). A multitask transformer for offensive language detection and target identification in hatebr. In *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 1049–1054.

Talat, Z. and Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93.