World Journal of Information Systems 2(4) 55-64
DOI: https://doi.org/10.17013/wjis.v2i4.51

Article

Detecting Hate Speech in Portuguese Social Media: A Deep Learning Approach for the Brazilian Presidential Election

Leila Weitzel[1,*], Thalessa Hungerbühler Daroz[1], Luan Pereira Cunha[1], Rafael Von Helde[1], Lucas Mendonça Morais[1]

[1] Universidade Federal Fluminense, Brazil

***** Correspondence: leila_weitzel@id.uff.br

(Article history: Received August 30, 2025; Received in revised form September 06, 2025; Accepted October 06, 2025; Available online December 15, 2025)

Abstract: Social media enables widespread communication with minimal access restrictions, facilitating the expression of emotions and opinions, including harmful content such as hate speech. The dissemination of offensive language online can negatively impact individuals and contribute to social instability. Recent advances in machine learning and natural language processing have improved automated detection of such content. This study focuses on hate speech detection in Portuguese during the 2018 Brazilian presidential elections. We evaluate different deep learning architectures, hyperparameters, and feature configurations to determine optimal parameters for hate speech mining. Additionally, we contribute a Portuguese hate speech lexicon and an annotated dataset, both publicly available on GitHub.

Keywords: Hate Speech; Machine Learning; Word Embeddings; Deep Learning.

1. Introduction

Social media platforms operate with minimal access restrictions, enabling widespread communication. Microblogging services such as Twitter (now X) and Weibo have become prominent global communication tools due to their immediacy and efficiency in information dissemination [1]. These platforms facilitate not only news sharing but also the exchange of diverse opinions and emotional expressions, including harmful content.

The pervasive use of online social media has shifted communication paradigms toward networked sociability, where interactions extend beyond traditional sender-receiver dynamics to include multiple actors—individuals, communities, and broader society [2]. While digital environments enhance information and knowledge exchange, they also amplify real-world conflicts. For instance, the rapid spread of misinformation during the COVID-19 pandemic—such as false claims about vaccines—demonstrated how unchecked expression

© 2025 ITMA. All rights reserved.

can directly harm public health [3]. Similarly, coordinated disinformation campaigns in political contexts, like the 2020 U.S. elections and Brazil's 2022 electoral process, have shown how social media can undermine democratic institutions [4].

Social media enables unfiltered expression, including offensive, derogatory, or inflammatory content. Recent cases, such as the rise in anti-Semitic and Islamophobic rhetoric following geopolitical conflicts (e.g., the Israel-Hamas war in 2023), illustrate how hate speech escalates real-world violence [5]. Likewise, the weaponization of free speech—through deepfakes, cyberbullying, or targeted harassment of marginalized groups—reveals systemic risks to social cohesion. While these platforms support democratic discourse, their lack of effective moderation creates opportunities for mass manipulation, defamation, and societal harm. Empirical studies link exposure to extremist content on platforms like YouTube and Telegram to radicalization trends in Europe and North America [6]. These examples underscore the tension between free expression and its societal consequences in digital spaces [7].

Based on the context described herein, this paper aims to classify hate speech, on Portuguese (PT-BR) language. Hate/offensive language identification is classification task in Natural Language Processing (NLP). It has been an active area of research in both academia and industry for the past two decades. Lately, many platforms of social media websites monitor user posts, thus, this leads to a pressing demand for methods to automatically identify suspicious posts. The vast majority of the hate content in circulation in Brazil is propagated by individuals (or groups) with extreme far-right sympathies. These kinds of comments have been most openly hostile and hateful towards the LGBT community, black/brown, indigenous and female community. These contents pose a huge challenge to societies, as they can lead to the radicalization of debates threatening democracy and the health of the population.

Societies need to develop adequate response mechanisms in order to find a balance between freedom of expression on one side and the ability to life without oppressive remarks without imposing rigid censorship regimes. Impacts of abusive language are detrimental, ranging from short-term emotional reactions (anger, fear, self-blame, etc.) to long-term psychological effects (low self-esteem, depression, etc.), causing mental and physical health issues (sleep disorder, headache, eating disorder, etc.) [7].

It must be stressed that the research focuses on Supervised Learning tasks (with binary classification). It is not the intention of the paper to delve into issues of linguistic theory or other Deep Learning approaches other than those discussed here.

Hate speech is broadly defined as language that promotes harm, discrimination, hostility, or violence. While often conflated with offensive language (e.g., profanity or slurs), this study focuses specifically on expressions of hate or degrading attitudes rather than general vulgarity. Hate speech detection falls under sentiment analysis, which employs four primary methodologies: Machine Learning (ML): Utilizes feature extraction (e.g., term frequency) from annotated corpora [8, 9]; Deep Learning (DL): Leverages neural networks to learn

© 2025 ITMA. All rights reserved.

patterns from labeled data [10-13]; Lexicon-Based: Relies on predefined lexicons to assign polarity to words or phrases [8, 14, 15].

2. Material and methods

This study utilizes a dataset of Portuguese-language tweets collected during the 2018-2019 Brazilian electoral period via Twitter's public Application Programming Interface (API) and the Tweepy Python library. At the time of collection, Twitter's API allowed unrestricted access to historical tweets for research purposes, enabling comprehensive data acquisition without cost or stringent usage limits. As of 2023, significant changes to Twitter's API policies (post-acquisition by X Corp.) have restricted free access to tweet datasets. The current tiered API pricing model and rate limits render large-scale data collection. The collected tweets were filtered for hate speech indicators using keyword-based sampling (e.g., racial slurs, political antagonism terms). To ensure relevance, we excluded non-Portuguese content and retweets. The final corpus was manually annotated by native speakers using guidelines adapted from [16], with inter-annotator agreement (Fleiss' $\kappa \geq 0.75$) validating label consistency.

Regarding the origin of the data the names of users or mentions have been omitted for privacy purposes. We collected 80,000 anonymized tweets (with usernames/mentions removed for privacy) using hate speech-related keywords. Examples included: "bolsobosta" (derogatory term combining the president's name with "dung"); "bolsominion" (pejorative term for far-right supporters);"petralha" (derogatory term for left-wing supporters) and "coxinha" (slang for conservative/reactionary groups). The dataset also included tweets from pro- and anti-government news accounts to ensure political balance. All identifying user information was omitted to protect privacy.

A lexicon named HALEX (HAte LExicon) with 512 terms of hate words (and terms) were created. We decide to remove profanity (swear words) terms from lexicon mostly because Brazilians have the habit of expressing themselves using swear words in both cases, either to express happiness or to express hatred, hence, we attempt to avoid the bias. The examples of hate terms are shown in Table I. It is important to note that some terms only make sense in the Portuguese language; the translation into English may seem meaningless.

2.1. Data preprocessing

One of the main difficulties in processing a tweet, besides being an unstructured sentence, is that most of sentence may not use the correct grammar. Tweets contain many typographical errors, individual instances of typos, phonetic substitution and others that cause lexical deviation. This can be even worse if our goal is to identify hate speech. Some individuals have the habit of masking hateful comments by inserting asterisks, spaces, or replacing characters with similar-sounding ones. For instance, the letter "a" is replaced by

© 2025 ITMA. All rights reserved.

the character @. Other issues are related to the use of acronym slang and fashion words, acronyms, internet shorthand [17].

Apart from the plain text, tweets can have others elements such as hashtags; hyperlinks (typically a bitly URL, i.e., a URL shortening service), emoticons and references to other users, as "mention" (@<user>). In order to tackle these issues, we normalize tweets with the following procedures: first we remove: mentions; emoji, URL, special characters ($, %, &, # etc.), digits and punctuation (full stops, commas etc.) and stopwords. Then we delete duplicates tweets and tweets starting with "RT" because they refer to a previous tweet. We normalize abbreviations of: Political Parties, Brazilian States and Portuguese internet shorthand (also called internet slang), e.g. "pq" replaced by "porquê" ("why"), "aki" - "aqui" ("here"); "hj" - "hoje" ("today") and etc. Repeated characters were also deleted ("odeiooooo" –hateeee replaced by "odeio". We manual split hashtags e.g. "lulatapreso" by "lula está preso" – lula is in jail.

After the processing steps described herein, we remove tweets with less than three tokens. In the context of Twitter, it is common to see shortened messages e.g.: "que bixa safada !!" (what fags dirty !!) that correspond to a hate sentence but can also correspond to a joke (particularly in Portuguese-BR). This tweet does not have a semantic interpretation for other people. It is basically understandable by the author (or the receiver) himself. It is common that the human annotator assigns it as hate speech, however, this supposition is based on beliefs of the annotator, and not on the real situation. Hence, afterward the complete processing step, we obtained 33,771 tweets. We perform undersampling technique in order to attempt the balance the class frequencies, thus a total of 12,700 tweets are remain (6,350 hate and 6,350 not hate). Three preprocessing configurations to assess their impact on model performance are performed:

- **Raw Data (B0):** Unprocessed tweets (n=12,700) served as our baseline to establish natural language patterns without modifications.
- **Fully Processed (B1):** Tweets (n=11,150) underwent complete preprocessing including: Text normalization and Stopwords removal. This configuration tested whether linguistic simplification improves detection accuracy.
- **Partially Processed (B2):** Tweets (n=10,575) received text normalization but retained stopwords, allowing evaluation of their potential semantic value in hate speech contexts.

This stratified approach enabled systematic comparison of preprocessing effects on classification performance, particularly important for Portuguese language processing where stopwords may carry contextual meaning in offensive expressions. The variation in sample sizes reflects natural data loss during cleaning procedures.

2.2. model selection and configuration

© 2025 ITMA. All rights reserved.

Neural networks are composed of interconnected layers that process input data to produce outputs. Key components include: Layer Types: Input layer, hidden layers, and output layer; Activation Functions: ReLU, SoftMax, etc.; Loss Functions: Cross-entropy, mean squared error, etc.; Optimizers: Adam, Nadam, RMSProp, SGD. We evaluated three deep neural network architectures: CNN (Convolutional Neural Network); LSTM (Long Short-Term Memory) and Bi-LSTM (Bidirectional LSTM).

For optimization, we compared Nadam (Adam with Nesterov momentum) and RMSProp. To establish performance benchmarks, we implemented:

Multinomial Naïve Bayes (MNB): A probabilistic classifier suitable for text data, even when features are non-numerical and Shallow Neural Network: A network with 1–2 hidden layers, providing insight into fundamental learning dynamics before scaling to deeper architectures. Baselines help identify model limitations and guide the selection of more complex architectures.

We conducted 36 experiments to systematically assess how model architecture and dataset features affect generalization in hate speech detection.

2.3. hyperparameters tuning

The hyperparameters tuning define how our model is actually structured. The process of setting the hyperparameters requires expertise and extensive Trial-Error. The Trial-Error is performed in a semi-automatic approach, where the hyperparameters are initialized with maximum and minimum values, and during training the performance was evaluated using a call-back (AUC and ROC-Curve). In order to find the optimal model architecture, we follow the steps as follow, we: define a model, define the range of possible values for all hyperparameters, define a method for sampling hyperparameter values, define evaluative criteria. The ranges tested were: batch size = [32, 64]; learn rate = [0.0001, 0.001]; optimizers: [Nadam, RMSProp, SGD]; Activation function: [ReLU (hidden layers), Sigmoid (output layer)]. To avoid overfitting, we use standard regularization techniques, e.g., early stopping (stops the training when a monitored quantity has stopped improving) and dropout. Dropout randomly select a neuron that will be ignored during training.

3. Main Results

The Table III shows the main outcomes achieved. One can be seen that there are not statistical differences between Bi-LSTM, CNN and LSTM except for LSTM-B0. This finding may show that the LSTM was affected by preprocessing step. As we expected, the ShallowNet achieved the worst performances. On the other hand, the MNB, surprisingly, showed a performance that can be considered acceptable (about ~88%).

© 2025 ITMA. All rights reserved.

Table 1: The accuracy and Loss Function results

Network	Dataset	Accuracy	Loss function
Bi-LSTM	B0	0,94	0,20
Bi-LSTM	B1	0,94	0,17
Bi-LSTM	B2	0,96	0,14
CNN	B0	0,94	0,17
CNN	B1	0,95	0,16
CNN	B2	0,96	0,15
LSTM	B0	0,69	0,63
LSTM	B1	0,91	0,20
LSTM	B2	0,95	0,15
MNB	B0	0,84	n/a
MNB	B1	0,87	n/a
MNB	B2	0,91	n/a
ShallowNet	B0	0,60	98,94
ShallowNet	B1	0,62	32,95
ShallowNet	B2	0,55	44,0

The table 1 presents accuracy and loss metrics for different neural networks and baseline models across three dataset variants (B0: raw, B1: fully preprocessed, B2: partially preprocessed). The table 2 shows the analysis of the main results.

Table 2: The analysis of the results

Model	Best Accuracy	Worst Accuracy	Key Observation
Bi-LSTM	0.96 (B2)	0.94 (B0/B1)	Consistent high performance across all datasets. Lowest loss (0.14).
CNN	0.96 (B2)	0.94 (B0)	Nearly matches Bi-LSTM, with slightly higher loss.
LSTM	0.95 (B2)	0.69 (B0)	**Highly preprocessing-dependent**; B0 performance poor.
MNB	0.91 (B2)	0.84 (B0)	Competitive for simpler models, but lacks loss metrics.
ShallowNet	0.62 (B1)	0.55 (B2)	Unstable (loss >30), unsuitable for this task.

Dataset Impact (B0 vs. B1 vs. B2) - Preprocessing Benefits:

© 2025 ITMA. All rights reserved.

- All models (except ShallowNet) improved with preprocessing (B1/B2 > B0).
- LSTM showed the largest gain: +0.26 accuracy (B0→B2).
- B2 (partial preprocessing) yielded best results for most models, suggesting stopwords may carry meaningful context in Portuguese hate speech.

Bi-LSTM/CNN Robustness:

- Minimal accuracy variation (≤0.02) across datasets, indicating resilience to preprocessing choices.
- Bi-LSTM and CNN are optimal for Portuguese hate speech detection, combining high accuracy (≥0.94) and low loss (≤0.20). LSTM is viable only with preprocessing (B1/B2).

Preprocessing Matters:

- Text normalization (B1/B2) is critical, especially for sequential models (LSTM).
- Stopword retention (B2) may preserve hate speech cues in Portuguese.

Baseline Limitations:

- MNB is decent for quick prototyping but lacks loss metrics for full evaluation.
- ShallowNet fails catastrophically (loss >30), ruling out shallow architectures.

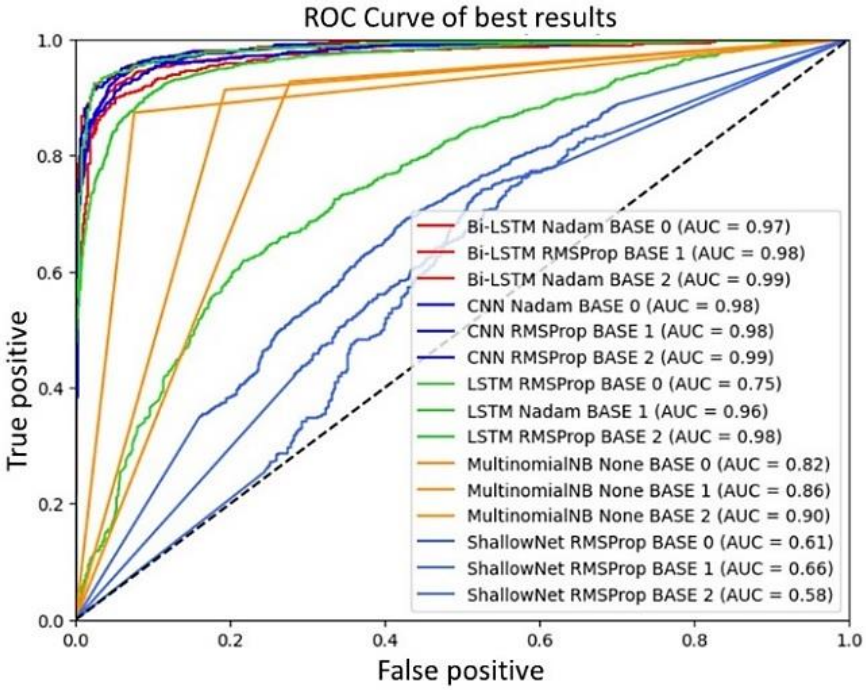

Figure 1. ROC of the best results

© 2025 ITMA. All rights reserved.

The ROC curves and AUC (Area Under the Curve), figure 1, values demonstrate the performance of different neural network architectures in classifying hate speech in Portuguese.

1. High-Performing Models (AUC ≥ 0.97)

- Bidirectional LSTM (Bi-LSTM): Achieved near-perfect discrimination (AUC = 0.97–0.99). Superior performance suggests its ability to capture contextual hate speech patterns in Portuguese text through bidirectional sequential analysis.
- LSTM (Medium): Consistently high AUC (0.99), indicating effectiveness in modeling temporal dependencies in hate speech.
- CNN + RMSProp: AUC = 0.99, showing that convolutional filters can extract meaningful hate speech features despite lacking sequential modeling.

2. Lower-Performing Models (AUC = 0.82)

- Shallow Networks Models: Limited discriminatory power (AUC = 0.82). Suggests that simplistic architectures fail to capture the complexity of hate speech in Portuguese. Baseline Limitations: Shallow networks and non-optimized models are inadequate for this task, reinforcing the need for deep learning in NLP hate speech applications.

Optimal Architectures: Bi-LSTM, LSTM, and CNN with RMSProp are best suited for Portuguese hate speech detection, as evidenced by near-perfect AUCs. The Context Matters, thus Sequential models (Bi-LSTM/LSTM) outperform non-sequential ones, highlighting the importance of linguistic context in hate speech.

4. Conclusion

The rise of social media as a primary platform for public discourse has amplified the spread of hate speech, necessitating robust detection mechanisms to mitigate its societal harm. This study conducted a comprehensive evaluation of deep learning models for hate speech identification in Brazilian Portuguese, analyzing their performance across different preprocessing strategies and architectural approaches. Our experiments revealed that bidirectional architectures (Bi-LSTM) and CNNs consistently achieved high accuracy (up to 0.96) and low loss (≤0.20), demonstrating their suitability for this task. Notably, these models performed robustly even with minimal preprocessing (B2), suggesting that stopword retention may preserve linguistically meaningful hate speech cues in Portuguese.

The LSTM model, while effective when applied to preprocessed data (B1/B2), exhibited significant performance drops on raw text (B0), underscoring the importance of text normalization for sequential models. In contrast, simpler approaches like Multinomial Naïve Bayes provided reasonable baseline accuracy (up to 0.91) but lacked the depth for reliable loss assessment, while shallow networks proved entirely inadequate due to catastrophic loss values (>30). These findings highlight the necessity of deep learning for capturing the contextual and morphological complexities of hate speech in Portuguese.

© 2025 ITMA. All rights reserved.

As contributions, we release a novel annotated dataset and a lexicon of 512 hate-related terms for Brazilian Portuguese, facilitating future research in this domain. All resources are publicly available to support reproducibility and further development.

For future work, we plan to address dataset limitations through oversampling techniques (e.g., SMOTE) to improve minority class representation and explore multi-label classification frameworks, which may better capture the nuanced spectrum of hate speech. Additionally, we aim to investigate hybrid architectures that combine the strengths of CNNs (local feature extraction) and Bi-LSTMs (contextual modeling) to further enhance detection capabilities. This research lays a foundation for more sophisticated and linguistically informed tools to combat online hate speech in Portuguese-speaking communities.

References

[1] Jairo L. Alves, Leila Weitzel, Paulo Quaresma, Carlos E. Cardoso, and Luan Cunha 2019. Brazilian Presidential Elections in the Era of Misinformation: A Machine Learning Approach to Analyse Fake News. Springer, City.

[2] Rousiley C. M. Maia, and Thaiane A. S. Rezende. 2016. Respect and Disrespect in Deliberation Across the Networked Media Environment: Examining Multiple Paths of Political Talk. 21, 2 %J J. Comp.-Med. Commun., https://doi.org/ 10.1111/jcc4.12155

[3] Sun Kyong Lee, Juhyung Sun, Seulki Jang, and Shane Connelly. 2022. Misinformation of COVID-19 vaccines and vaccine hesitancy. Scientific Reports 12, 1, https://doi.org/ 10.1038/s41598-022-17430-6

[4] Jaren Haber, Lisa Singh, Ceren Budak, Josh Pasek, Meena Balan, Ryan Callahan, Rob Churchill, Brandon Herren, and Kornraphop Kawintiranon. 2021. Research note: Lies and presidential debates: How political misinformation spread across media streams during the 2020 election. HKS Misinfo Review https://doi.org/ 10.37016/mr-2020-84

[5] Global threats to freedom of expression arising from the conflict in Gaza - Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Irene Khan (A/79/319). City.

[6] Abdullah Alrhmoun, Charlie Winter, and János Kertész. 2024. Automating Terror: The Role and Impact of Telegram Bots in the Islamic State's Online Ecosystem. Terrorism and Political Violence 36, 4, https://doi.org/ 10.1080/09546553.2023.2169141

[7] Thiago Chu, Leila Weitzel, and Paulo Quaresma 2024. Comparative Analysis of Hate Speech Detection Models on Brazilian Portuguese Data: Modified BERT vs. BERT vs. Standard Machine Learning Algorithms. SCITEPRESS - Science and Technology Publications, City.

[8] Gabriel Nascimento, Flavio Carvalho, Alexandre Martins da Cunha, Carlos Roberto Viana, and Gustavo Paiva Guedes 2019. Hate speech detection using Brazilian imageboards. City.

[9] Adriano Silva, and Norton Roman 2020. Hate Speech Detection in Portuguese with Naïve Bayes, SVM, MLP and Logistic Regression. Sociedade Brasileira de Computação - SBC, City.

[10] João Augusto Leite, Diego Silva, Kalina Bontcheva, and Carolina Scarton 2020. Toxic Language Detection in Social Media for Brazilian Portuguese: New Dataset and Multilingual Analysis. Association for Computational Linguistics, City.

[11] Felix L. V. da Silva, and Larissa Astrogildo de Freitas. 2022. Brazilian Portuguese Hate Speech Classification using BERTimbau. The International FLAIRS Conference Proceedings

© 2025 ITMA. All rights reserved.

[12] L. Weitzel, T. H. Daroz, L. P. Cunha, R. V. Helde, and L. M. d. Morais 2023. Investigating Deep Learning Approaches for Hate Speech Detection in Social Media : Portuguese-BR tweets. IEEE xplore, City.

[13] Oliveira Felipe Ramos De, Reis Victoria Dias, and Ebecken Nelson Francisco Favilla. 2024. Detecting Hate Speech on Brazilian Social Media: New Dataset and Analysis. CILAMCE https://doi.org/ 10.55592/cilamce.v6i06.8208

[14] Rogers Pelle, and Viviane Pereira Moreira 2017. Offensive Comments in the Brazilian Web: a dataset and baseline results. City.

[15] Francielle Vargas, Isabelle Carvalho, Fabiana Rodrigues de Góes, Thiago Pardo, and Fabrício Benevenuto 2022. HateBR: A Large Expert Annotated Corpus of Brazilian Instagram Comments for Offensive Language and Hate Speech Detection. European Language Resources Association, City.

[16] Joseph Fleiss, Bruce Levin, and Myunghee Paik. 2003. In Statistical Methods for Rates and Proportions. Statistical Methods for Rates and Proportions 203https://doi.org/ 10.1002/0471445428

[17] H. Watanabe, M. Bouazizi, and T. Ohtsuki. 2018. Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection. IEEE Access 6https://doi.org/ 10.1109/ACCESS.2018.2806394

© 2025 ITMA. All rights reserved.