XLVI CILAMCE
VITÓRIA - ES - BRAZIL - 2025
XLVI IBERO-LATIN AMERICAN CONGRESS ON COMPUTATIONAL METHODS IN ENGENEERING

Detecting Misogynistic Hate Speech in Portuguese

Ana Rúbia Ramos Vicente[1], Kelly Assis de Souza Gazolli[1], Karin Satie Komati[1]

[1]*Graduate program in Applied Computing (PPComp)*
Instituto Federal do Espírito Santo (IFES), Campus Serra
Av. dos Sabiás, 330 - Morada de Laranjeiras, 29166-630, Serra-ES, Brazil
anarubiarvicente@gmail.com, {kasouza, kkomati}@ifes.edu.br

Abstract. Misogynistic hate speech targets women with language that promotes discrimination and hostility. In Brazil, the scarcity of annotated datasets in Portuguese presents additional challenges for automated detection. This study evaluates supervised machine learning models for classifying misogynistic hate speech using the MINA-BR dataset, which contains over 2,000 manually labeled comments. Four experimental scenarios were considered: (1) Bag of Words, (2) TF-IDF (Term Frequency-Inverse Document Frequency) with unigrams, (3) TF-IDF with bigrams, and (4) TF-IDF with SMOTE (Synthetic Minority Over-sampling Technique) for class balancing. Text preprocessing included lemmatization and stopword removal. Feature vectors were used as input for six classifiers: Naive Bayes (NB), Support Vector Machines (SVM), Logistic Regression (LR), Multilayer Perceptron (MLP), Random Forest (RF), and XGBoost. Model evaluation was performed using 5-fold cross-validation to ensure robustness and generalization of results. Among the models tested, Random Forest with TF-IDF and SMOTE representation achieved the highest accuracy on the test set, reaching 86.8%. However, the best F1-score was obtained by LR with TF-IDF and SMOTE, achieving a score of 0.565, indicating a more balanced performance between precision and recall. Additionally, the use of SMOTE notably improved recall in models such as NB and LR, enhancing the detection of the minority class ("Hate"). These results highlight the importance of combining appropriate text representation techniques with class-balancing strategies to develop effective hate speech detection systems in Portuguese-language content.

Keywords: Natural Language Processing, Text Classification, imbalanced dataset.

1 Introduction

Hate speech refers to any form of expression—spoken, written, or behavioral—that targets individuals or groups using derogatory or discriminatory language based on aspects of their identity [1]. Misogynistic hate speech, specifically, is directed at women or girls and is characterized by expressions that reinforce gender-based discrimination, hostility, or contempt [2]. With the expansion of global digital communication, hate speech on social media platforms has become increasingly frequent, highlighting the need for effective automated tools for its detection and mitigation. In Brazil, this challenge is compounded by the scarcity of annotated datasets in Portuguese. One initiative that seeks to fill this gap is MINA-BR, a publicly available Brazilian Portuguese dataset composed of internet comments containing hate speech against women [3, 4]. Over 2,000 of these comments were manually annotated by three independent reviewers.

This study aims to compare artificial intelligence algorithms in the task of misogynistic hate speech detection in Portuguese, using the MINA-BR corpus. The methodology comprises four experimental scenarios: (1) Bag of Words (BoW), (2) TF-IDF (Term Frequency–Inverse Document Frequency) with unigrams, (3) TF-IDF with bigrams, and (4) TF-IDF with the application of SMOTE (Synthetic Minority Over-sampling Technique) for class balancing.

In all scenarios, the comments undergo a text preprocessing step that includes lemmatization and removal of stopwords. One of the main differences between this study and the original work that introduced the MINA-BR corpus lies in the text preprocessing stage. While the previous study employed the stemming technique for word normalization, this study opts for lemmatization. Lemmatization takes into account the context and morphology of words, allowing terms to be reduced to their canonical form in a linguistically and semantically consistent manner. This approach tends to yield more accurate results in natural language processing tasks that require semantic understanding, such as hate speech detection. In contrast, although stemming is computationally simpler

and faster, it may produce truncated word stems that do not correspond to real words in the language, potentially compromising text representation and, consequently, classification model performance [5].

The vector representations obtained are used as input for widely adopted supervised learning models in the literature, such as Naive Bayes (NB), SVM (Support Vector Machines), Logistic Regression (LR), Multilayer Perceptron (MLP), Random Forest (RF), and XGBoost. Given the imbalanced nature of the dataset, the study also investigates resampling strategies, with emphasis on the use of SMOTE, aiming to improve the models' ability to correctly identify the minority class ("Hate"). Additionally, the impact of different vectorization and balancing techniques on classifier performance is analyzed. Model evaluation was performed using 5-fold cross-validation to ensure robustness and generalization of results [6]. To evaluate the results, the metrics of precision, recall, and F1-score are used. The following sections are organized as follows: Section II presents related work; Section III describes the dataset and the methods used; Section IV discusses the results obtained; and finally, Section V presents the final considerations.

2 Related Work

The work by Plath et al. [4] describes the construction process of the MINA-BR dataset, in Brazilian Portuguese, focusing on the identification of hate speech against women. The study also presented experiments using text mining techniques and classification algorithms. The preprocessing steps applied included: (i) conversion to lowercase; (ii) removal of hashtags and duplicate white spaces; (iii) sentence tokenization; (iv) removal of stopwords based on the Portuguese list provided by the NLTK library; and (v) stemming, a technique that removes suffixes or prefixes to obtain the "root" of a word (e.g., "correndo", "correu", "corre" $\rightarrow$ "corr"). Preprocessing was evaluated in two variations: with and without stemming. For text representation, the BoW and TF-IDF models were used. The classifiers evaluated included linear SVM, RF, and three variants of NB (Multinomial, Bernoulli, and Gaussian). The best result was achieved by the SVM classifier with TF-IDF, without the application of stemming, reaching an F1-score of 0.57.

This study by Rupapara et al. [7] addresses the critical problem of toxic comment classification on social media platforms, particularly focusing on the challenges posed by imbalanced datasets. The authors propose a novel ensemble approach called Regression Vector Voting Classifier (RVVC) , which combines LR and SVM under soft voting criteria. To counter data imbalance, the study extensively evaluates SMOTE and random under-sampling, alongside feature extraction methods such as TF-IDF and BoW. Experimental results demonstrate that data balancing significantly increases classification accuracy, with models performing poorly on imbalanced datasets. RVVC consistently outperforms individual machine learning models and other state-of-the-art approaches, achieving its highest accuracy of 0.97 when utilizing TF-IDF features on a SMOTE over-sampled dataset. The research highlights that over-sampling leads to better performance than under-sampling due to an increased feature set, and TF-IDF generally provides superior classification accuracy compared to BoW.

The paper of Vargas et al. [8] addresses the data scarcity in low-resource hate speech detection by introducing two expert-annotated resources: the HateBR 2.0 corpus (7,000 Brazilian Portuguese Instagram comments) and the Multilingual Offensive Lexicon (MOL, 1,000 terms with contextual information). Both achieved high inter-annotator agreement, underscoring the importance of expert-annotated, context-aware data. For baseline experiments, the study employed feature representation models like BoW, TF-IDF, fastText, and mBERT, alongside learning algorithms such as NB, SVM, fastText, and mBERT. Models trained on HateBR 2.0 achieved an average F1-score of 85%, demonstrating that the new corpus provides more reliable benchmarks than previous Portuguese hate speech datasets.

3 MinaBR

In this study, the MINA-BR dataset, developed by the University of Campinas (Unicamp), was employed [3, 4]. The main goal of the MINA-BR dataset is to support the development of applications aimed at combating violence against women in Brazilian Portuguese. Specifically, the labeled version of the dataset, called "MINA-BR Rotulada," was used. It contains 2,091 comments extracted from the social media platforms YouTube and Twitter. The comments were collected using keywords selected from an online survey conducted with Unicamp students. In this work, only two fields were used for each comment: (i) "comment," which is the full text of the comment being analyzed, and (ii) the binary classification field "odio," with HATE (1) and NON-HATE (0) as possible values. Out of the total 2,091 comments, 1,751 were labeled as "Non-Hate" (0) and 340 as "Hate" (1).

By adopting a collaborative labeling approach, the project faced a data poisoning attack during the collection phase, resulting in the insertion of false data, especially regarding the hate speech classification. Although the dataset already exhibited an inherently imbalanced nature—a common characteristic of hate speech corpora [1,

9]—the attack further exacerbated this condition. Subsequent analyses revealed the creation of bots designed to simulate social profiles of young, highly educated women in order to manipulate the survey and, consequently, the labels. As a result of this incident, collaborative labeling via website was discontinued. Nevertheless, the team remains interested in resuming this methodology in the future to expand and enrich the dataset. It is important to note that the data poisoning attack was not the only form of opposition encountered by the project, which highlights the need for greater visibility and support for initiatives addressing online violence against women.

4 Methods

In this work, the preprocessing of the comments (the "comment" column) from the MINA-BR dataset included the following steps: (i) conversion to lowercase; (ii) removal of non-alphanumeric characters; (iii) tokenization; (iv) lemmatization; and (v) removal of Portuguese stopwords based on the list provided by the NLTK library. Lemmatization was performed using NLTK's `WordNetLemmatizer`, reducing each token to its base form (lemma), which allows different inflections of the same word to be treated as a single term (e.g., "correndo" $\rightarrow$ "correr").

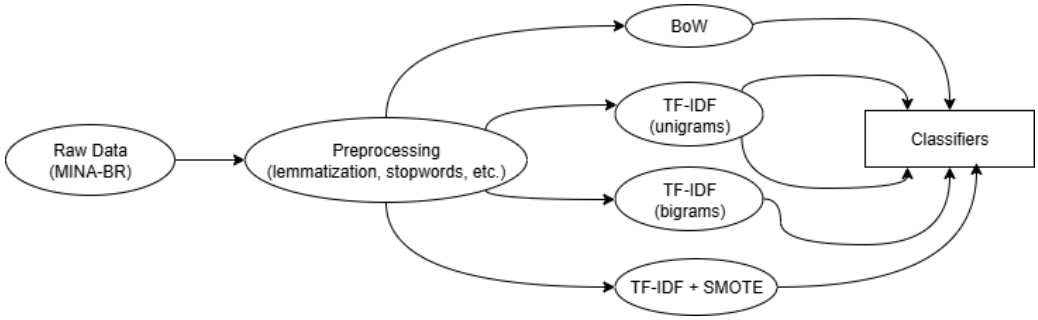


Figure 1. Flowchart of the Four Experimental Scenarios for Hate Speech Detection.

For vectorization, which converts text into numerical format, two techniques were employed: BoW and TF-IDF. BoW represents documents as collections of words without considering word order, focusing on the presence and frequency of terms. TF-IDF, on the other hand, weights terms based on their frequency in a document and their rarity across the entire corpus, allowing the model to capture the relative importance of each word [10]. Figure 1 presents the flowchart with the four experimental scenarios evaluated: (1) BoW, (2) TF-IDF with unigrams, (3) TF-IDF with bigrams, and (4) TF-IDF with SMOTE applied for minority class balancing.

In the classification stage, we employed six widely used supervised learning algorithms from the literature: NB, linear SVM, LR, MLP, RF, and XGBoost. NB is based on Bayes' theorem, assuming conditional independence among input variables, and is particularly efficient in high-dimensional scenarios. SVMs aim to maximize the margin between classes by constructing hyperplanes with strong generalization capability for binary classification tasks. LR models the probability of class occurrence through a linear combination of input variables, being highly interpretable and effective in linear problems. MLP is a feedforward neural network composed of one or more hidden layers, capable of capturing complex nonlinear relationships between attributes. RF is an ensemble method based on building multiple decision trees trained on random subsets of data and features, offering robustness against overfitting. Finally, XGBoost is a gradient boosting algorithm that builds additive models through sequentially optimized decision trees, known for its high performance in classification tasks involving structured data.

The trained models were evaluated using 5-fold cross-validation (CV). In this method, the dataset is randomly divided into five equally sized subsets, or folds. For each iteration, one fold is used for testing while the remaining four folds are used for training. The process is repeated until each fold has served once as the test set. As stated by [11], experimental results have shown that the k-fold cross-validation technique is particularly suitable for handling imbalanced datasets, providing more stable and reliable performance estimates.

5 Results and Discussion

The MINA-BR dataset, loaded from the file `MinaBR.csv`, was split into training (80%) and testing (20%) sets, selected randomly. The performance metrics collected are: precision, recall, F1-score, and accuracy, calculated based on the final classification of the positive class (HATE).

5.1 Scenario 1: Bag of Words

In the first experimental scenario, the preprocessed textual data were vectorized using the BoW technique. The results obtained by each classifier are presented in Table 1 and loss curve in Figure 2.

Table 1. Results of the models using Bag of Words (CV K=5)

Model	**Precision**	**Recall**	**F1-Score**	**Accuracy**
Naive Bayes	**0.658**	0.282	0.395	0.859
SVM	0.472	**0.438**	0.454	0.829
MLP	0.494	0.379	0.428	0.836
Logistic Regression	0.586	0.359	0.444	0.854
Random Forest	0.643	0.350	0.453	**0.863**
XGBoost	0.568	0.424	**0.483**	0.853

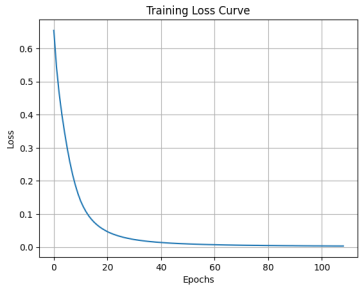


Figure 2. Training loss curve of MLP - Scenario 1

The results indicate that the Naive Bayes classifier achieved a precision of 0.658, meaning that most of the instances it classified as hate speech were correct. However, its recall was low (0.282), indicating that the model failed to identify the majority of the positive cases. XGBoost achieves the best F1-score (0.483) with competitive accuracy (0.853), while LR offers a balanced profile (precision 0.586, recall 0.359, F1 0.444, accuracy 0.854). RF delivers the best accuracy (0.863) with good precision (0.643) but moderate recall (0.350), resulting in an F1 of 0.453; MLP remains intermediate across metrics (F1 0.428, accuracy 0.836). Overall, the results reveal a clear precision–recall trade-off: NB prioritizes precision, SVM emphasizes recall, while XGBoost achieves the best F1-score and RF attains the highest accuracy.

5.2 Scenario 2: TF-IDF

In this second scenario, the preprocessing methodology was maintained, and the textual data were vectorized using the TF-IDF technique. The results obtained by the classifiers are presented in Table 2 and loss curve in Figure 3.

Table 2. Results of the models using TF-IDF (CV K=5)

Model	**Precision**	**Recall**	**F1-Score**	**Accuracy**
Naive Bayes	**0.713**	0.044	0.082	0.841
SVM	0.630	0.376	0.468	**0.862**
MLP	0.458	0.344	0.393	0.827
Logistic Regression	0.695	0.241	0.353	0.857
Random Forest	0.672	0.297	0.406	0.861
XGBoost	0.544	**0.429**	**0.479**	0.848

With TF-IDF, NB attains the highest precision (0.713) but an extremely low recall (0.044), meaning it flags few positive cases despite being very accurate when it does. SVM delivers the best accuracy (0.862) and a competitive F1-score (0.468), offering a more balanced trade-off than NB. XGBoost achieves the top F1-score (0.479) and the highest recall (0.429), favoring the identification of more positive instances at the cost of lower precision

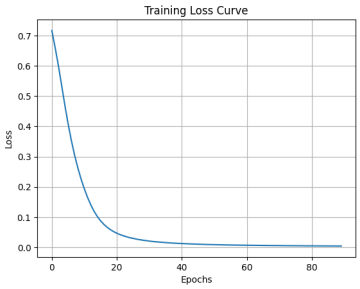


Figure 3. Training loss curve of MLP - Scenario 2

(0.544). RF posts near-best accuracy (0.861) and strong precision (0.672), but its recall is modest (0.297), yielding an F1 of 0.406. LR is also precision-oriented (0.695) with low recall (0.241), leading to an F1 of 0.353 and accuracy of 0.857. The MLP trails the others (F1 0.393; accuracy 0.827). Overall, TF-IDF accentuates a clear precision–recall trade-off: XGBoost is preferable when recall/F1 is prioritized, whereas SVM, RF, and LR provide higher accuracy or precision depending on the objective

5.3 Scenario 3: TF-IDF with N-grams (Bigrams)

In this third scenario, the TF-IDF vectorization was adjusted to include bigram n-grams (pairs of words), in addition to individual words. This feature allows the models to consider local relationships between adjacent terms in the preprocessed text, which includes lemmatization and stopword removal. The results obtained by the classifiers are presented in Table 3 and loss curve in Figure 4.

Table 3. Results of the models using TF-IDF with Bigrams (CV K=5)

Model	**Precision**	**Recall**	**F1-Score**	**Accuracy**
Naive Bayes	**0.880**	0.021	0.040	0.839
SVM	0.604	**0.374**	**0.458**	**0.857**
MLP	0.554	**0.374**	0.441	0.848
Logistic Regression	0.501	0.247	0.328	0.837
Random Forest	0.655	0.256	0.365	0.856
XGBoost	0.590	0.335	0.423	0.854

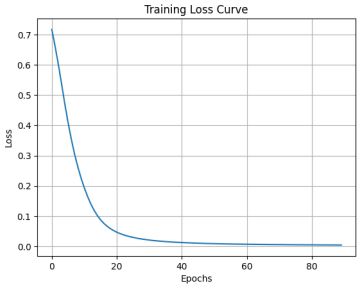


Figure 4. Training loss curve of MLP - Scenario 3

With the inclusion of bigrams, NB reinforced a high precision (0.880) but low recall (0.021), yielding an F1 of 0.040. When it flags hate speech, it is usually right, but it rarely flags it. SVM delivered the best F1 (0.458) and the highest accuracy (0.857), offering the most competitive balance between precision (0.604) and recall (0.374). MLP tied for the highest recall (0.374) and reached F1 = 0.441 (accuracy 0.848), suggesting bigrams benefited this model more than most others. RF (F1 = 0.365; accuracy 0.856) and XGBoost (F1 = 0.423; accuracy 0.854) maintained high accuracies with modest F1 gains, while LR lagged on F1 (0.328), indicating that the bigram representation likely increased sparsity or introduced features with lower discriminative value for this classifier.

5.4 Scenario 4: Balancing with SMOTE and TF-IDF

To address the class imbalance present in the MINA-BR dataset, this scenario applied the SMOTE technique to the training set. SMOTE generated synthetic instances of the minority class ("hate") with the aim of balancing the class distribution before training the models. Text vectorization was performed using the TF-IDF technique, without the inclusion of n-grams, and SMOTE was applied after this vectorization step. The models were trained on the balanced dataset and evaluated using the original test set. The results obtained are presented in Table 4 and loss curve in Figure 5.

Table 4. Results of the models using TF-IDF with SMOTE (CV K=5)

Model	**Precision**	**Recall**	**F1-Score**	**Accuracy**
Naive Bayes	0.375	**0.685**	0.484	0.762
SVM	0.505	0.544	0.523	0.839
MLP	0.439	0.368	0.400	0.821
Logistic Regression	0.522	0.618	**0.565**	0.846
Random Forest	**0.617**	0.497	0.549	**0.868**
XGBoost	0.516	0.494	0.504	0.842

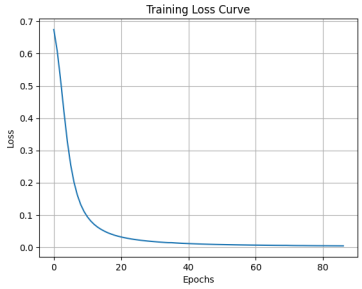


Figure 5. Training loss curve of MLP - Scenario 4

Applying SMOTE chiefly boosts recall across models. NB attains the highest recall (0.685) but at the cost of precision (0.375), yielding an F1 of 0.484. LR offers the best overall balance with the top F1 (0.565) and strong recall (0.618), alongside an accuracy of 0.846. RF delivers the highest precision (0.617) and best accuracy (0.868), with a solid F1 of 0.549, indicating reliable positives while capturing a reasonable share of the minority class. SVM remains well-balanced (precision 0.505, recall 0.544, F1 0.523; accuracy 0.839), and XGBoost shows moderate gains (F1 0.504). MLP trails in F1 (0.400) due to lower recall (0.368). Overall, class balancing via SMOTE helps models learn minority-class patterns, improving recall and, for LR and RF in particular, leading to stronger F1 and competitive accuracy.

6 Conclusions

This ongoing study conducted a comparative analysis of several machine learning models for hate speech classification, using the MINA-BR dataset, which comprises two main classes: "Hate" and "Non-Hate." The F1-score was adopted as the primary performance metric in this evaluation, as it offers a balanced view of precision and recall—crucial in imbalanced classification tasks such as hate speech detection.

Among all configurations, Logistic Regression with TF-IDF + SMOTE delivers the highest F1-score (0.565), offering the most balanced performance for the minority class. The highest accuracy is achieved by Random Forest with TF-IDF + SMOTE (0.868), though its recall is moderate (0.497). Without SMOTE, the best F1 varies by representation—XGBoost with Bag-of-Words (0.483) and SVM with TF-IDF + bigrams (0.458)—but both trail the SMOTE-enhanced LR. Naive Bayes with TF-IDF + bigrams attains the highest precision (0.880) yet exhibits extremely low recall (0.021), making it unsuitable when coverage of hate speech is a priority. Overall, SMOTE consistently boosts recall—notably for NB (to 0.685) and LR (to 0.618)—and raises F1, underscoring the importance of class balancing for this imbalanced dataset.

It is worth noting that, unlike the original study introducing the MINA-BR dataset—which adopted stemming as part of the preprocessing pipeline—this study employed lemmatization. While lemmatization is theoretically advantageous for preserving the canonical forms of words, the results observed across all experimental scenarios

suggest that its impact on classification performance was limited. No consistent improvement in F1-score was observed when compared to prior results using stemming. This indicates that, for this specific task and dataset, the choice between lemmatization and stemming may not be a decisive factor in model performance, especially when simpler term-weighting schemes such as TF-IDF are used.

As future work, we propose the integration of pre-trained language models, such as BERT (Bidirectional Encoder Representations from Transformers), which leverage deep contextual embeddings to capture semantic and syntactic nuances more effectively. Prior studies, such as the work by Saleh, Alhothali, and Moria [12], have shown that BERT-based architectures outperform traditional models in hate speech detection tasks. In addition to exploring BERT, we intend to investigate multilingual variants (e.g., mBERT or XLM-R) to enable broader applicability across dialects and regional variations of Portuguese. Another line of future work includes the expansion of the dataset through data augmentation techniques, such as back-translation or paraphrasing, to mitigate class imbalance and improve generalization. Finally, we aim to explore ensemble strategies that combine classical machine learning models with transformer-based approaches, potentially enhancing both performance and robustness.

Acknowledgements. The authors would like to thank IFES (EDITAL PRPPG 08/2025). Professor Komati thanks CNPq for the DT-2 grant (nº 302726/2023-3) and project nº407742/2022-0, also thanks FAPES for project nº 1023/2022 P:2022-8TZV6.

Authorship statement. The authors hereby confirm that they are the sole liable persons responsible for the authorship of this work, and that all material that has been herein included as part of the present paper is either the property (and authorship) of the authors, or has the permission of the owners to be included here.

References

[1] de R. Pelle and V. Moreira. Offensive comments in the Brazilian web: a dataset and baseline results. In *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*, pp. 510–519, Porto Alegre, RS, Brasil. SBC, 2017.

[2] P. Saha, B. Mathew, P. Goyal, and A. Mukherjee. Hateminers: Detecting hate speech against women. *arXiv preprint arXiv:1812.06700*, 2018.

[3] H. d. O. Plath, de M. E. Oliveira Paiva, D. L. Pinto, and P. D. P. Costa. MINA-BR: base de dados para identificação de discurso de ódio contra a mulher na internet, 2025.

[4] H. O. Plath, M. E. O. Paiva, D. L. Pinto, and P. D. P. Costa. Detecção de discurso de Ódio contra mulheres em textos em português brasileiro: Construção da base MINA-BR e modelo de classificação. *Revista Eletrônica de Iniciação Científica em Computação*, vol. 20, n. 3, 2022.

[5] R. Pramana, J. J. Subroto, A. A. S. Gunawan, and others. Systematic literature review of stemming and lemmatization performance for sentence similarity. In *2022 IEEE 7th international conference on information technology and digital applications (ICITDA)*, pp. 1–6. IEEE, 2022.

[6] S. K. Mohapatra, S. Prasad, D. K. Bebarta, T. K. Das, K. Srinivasan, and Y.-C. Hu. Automatic hate speech detection in english-odia code mixed social media data using machine learning techniques. *Applied Sciences*, vol. 11, n. 18, pp. 8575, 2021.

[7] V. Rupapara, F. Rustam, H. F. Shahzad, A. Mehmood, I. Ashraf, and G. S. Choi. Impact of SMOTE on imbalanced text features for toxic comments classification using RVVC model. *IEEE Access*, vol. 9, pp. 78621–78634, 2021.

[8] F. Vargas, I. Carvalho, T. A. Pardo, and F. Benevenuto. Context-aware and expert data resources for Brazilian Portuguese hate speech detection. *Natural Language Processing*, vol. 31, n. 2, pp. 435–456, 2025.

[9] de O. Gibert, N. Perez, A. García-Pablos, and M. Cuadros. Hate speech dataset from a white supremacy forum. In D. Fišer, R. Huang, V. Prabhakaran, R. Voigt, Z. Waseem, and J. Wernimont, eds, *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pp. 11–20, Brussels, Belgium. Association for Computational Linguistics, 2018.

[10] I. A. Tyosar Lubem. Stephen Akuma. Comparing bag of words and TF-IDF with different models for hate speech detection from live tweets. *Miscellaneous*, vol. 14, n. 7, pp. 3629–3635, 2022.

[11] P. K. Roy, A. K. Tripathy, T. K. Das, and X.-Z. Gao. A framework for hate speech detection using deep convolutional neural network. *IEEE Access*, vol. 8, pp. 204951–204962, 2020.

[12] H. Saleh, A. Alhothali, and K. Moria. Detection of hate speech using BERT and hate speech word embedding with deep model. *Applied Artificial Intelligence*, vol. 37, n. 1, pp. 2166719, 2023.