Detecting Hate Speech on Brazilian Social Media: New Dataset and Analysis

Felipe R. Oliveira[1], Victoria D. Reis[1], Nelson F. F. Ebecken[1]

[1]*Dept. of Civil Engineering,, Federal University of Rio de Janeiro*
Pedro Calmon, 550, 20210-030, Rio de Janeiro, Brazil
Felipe.oliveira@coc.ufrj.br

Abstract. Social media plays a crucial role in human interaction, facilitating communication and self-expression. However, the proliferation of hate speech on these platforms poses significant risks to individuals and communities. Detecting and addressing hate speech is particularly challenging in languages like Portuguese due to its rich vocabulary, complex grammar, and regional variations. To address this challenge, we introduce TuPy-E, the largest annotated Portuguese corpus dedicated to hate speech detection. Through a comprehensive analysis utilizing advanced techniques such as BERT and GPT-2 models, our research contributes to both academic understanding and practical applications in this field.

Keywords: hate speech, classification, social media, natural processing language.

1. Introduction

Online hate speech shares fundamental similarities with its offline counterpart but is distinguished by its unique interactions and use of specific vocabulary, accusations, and conspiracy theories that can emerge, proliferate, and disappear rapidly. These messages have the ability to go viral within an extremely short period, often within minutes.

The spread of online hate speech, acknowledged by the UN Special Rapporteur on Minority Issues of the Human Rights Council, presents distinct challenges. Both social media platforms and organizations dedicated to combating hate speech recognize a significant increase in the prevalence of these messages on the internet, demanding unprecedented attention to develop appropriate responses.

According to HateBase, an online application that catalogues instances of hate speech worldwide, most instances of hate speech target individuals based on ethnicity and nationality, although incitements related to religion and social class have also grown (Davidson et al., 2017) [1].

While online hate speech is not fundamentally different from its offline counterpart, it presents specific and unique challenges in terms of content and regulation. These challenges are linked to the persistence of content, its dissemination, the anonymity of perpetrators, and the complexity of crossing jurisdictional borders (Benesch, 2021; Gagliardone et al., 2015; UNESCO, 2021; Wu et al., 2022) [2,3,4,5].

One significant barrier in automated hate speech detection lies in the scarcity of publicly available and properly annotated datasets, with the majority focused on English. A survey conducted by Jahan & Oussalah (2023) [6], listing major works published over the last decade, highlights English as present in 51% of the datasets, with Portuguese representing only 1% of the compilation identified by the authors.

In the literature, there is a shortage of publicly available tools utilizing automatic hate speech detection techniques. One goal of this work is to provide an open-source tool. Additionally, there is a notable scarcity of collected and annotated datasets, preferably organized by hate speech category (such as racism, xenophobia, among others).

Recognizing the importance of previous research in this domain and the absence of annotated datasets for automated hate speech detection in Portuguese, we propose TuPy Expanded, or TuPy-E. This initiative aims to consolidate the original dataset presented in this research (TuPy) with the findings of Fortuna et al. (2019), Leite et al. (2020), and Vargas et al. (2022) [7, 8, 9, 10].

Based on this, we propose refining various models for the two classification tasks using the dataset developed in this dissertation: binary classification and categorization of hate speech in Portuguese. The approach involves leveraging three refined language models - BERT-Base, BERT-Large, and GPT-2 Small.

2. Dataset

To address the significant gaps in current Portuguese hate speech repositories, we introduce the TuPy-E dataset. This dataset builds on earlier research and the lack of annotated data for automated hate speech detection by combining insights from Fortuna et al. (2019), Leite et al. (2020), and Vargas et al. (2022) with a new, proprietary dataset.

For the unpublished segment of the TuPy-E dataset, we dedicated approximately seven months—from March 2023 to September 2023—to constructing the corpus. This effort involved collaboration with a multidisciplinary team, including a linguist, a human rights lawyer, several behavior psychologists with master's degrees, and experts in NLP and machine learning.

Our approach followed a framework inspired by Vargas et al. (2022) and Fortuna (2017), applying rigorous criteria for selecting annotators. The criteria included:

i) A range of political viewpoints, including right-wing, liberal, and far-left perspectives.

ii) Advanced academic qualifications, involving individuals with master's degrees, doctoral candidates, and PhD holders.

iii) Specialization in fields relevant to the focus and goals of our research.

To integrate data from key studies in automatic hate speech detection in Portuguese, we created a unified database by combining labeled document sets from Fortuna et al. (2019), Leite et al. (2020), and Vargas et al. (2022). To ensure the coherence and compatibility of our dataset, we followed these integration guidelines:

i) Fortuna et al. (2019) developed a database with 5,670 tweets, each labeled by three separate annotators to identify hate speech. To ensure consistency, we utilized a majority-voting method for classifying these documents.

ii) The dataset from Leite et al. (2020) includes 21,000 tweets labeled by 129 volunteers, with each tweet assessed by three different evaluators. This dataset covers six types of toxic speech: homophobia, racism, xenophobia, offensive language, obscene language, and misogyny. Tweets with offensive and obscene language were excluded from the hate speech categorization. We also applied a majority-voting process for classification based on these criteria.

iii) Vargas et al. (2022) compiled a set of 7,000 Instagram comments, labeled by three annotators. These comments had already undergone a majority-voting process, so no further classification was needed.

After these integration steps, the corpus was annotated at two levels. The first level involved a binary classification to differentiate between aggressive and non-aggressive language. In the second level, we categorized each tweet marked as aggressive into specific hate speech categories, including ageism, aporophobia, body shaming, capacitism, LGBTphobia, political hate, racism, religious intolerance, misogyny, and xenophobia. It is important to note that a single tweet could belong to one or more of these categories. For more information on the methodology for creating the proprietary dataset, see Oliveira (2024) [14]. The subsequent diagram presents the process schematics (Figure 1).

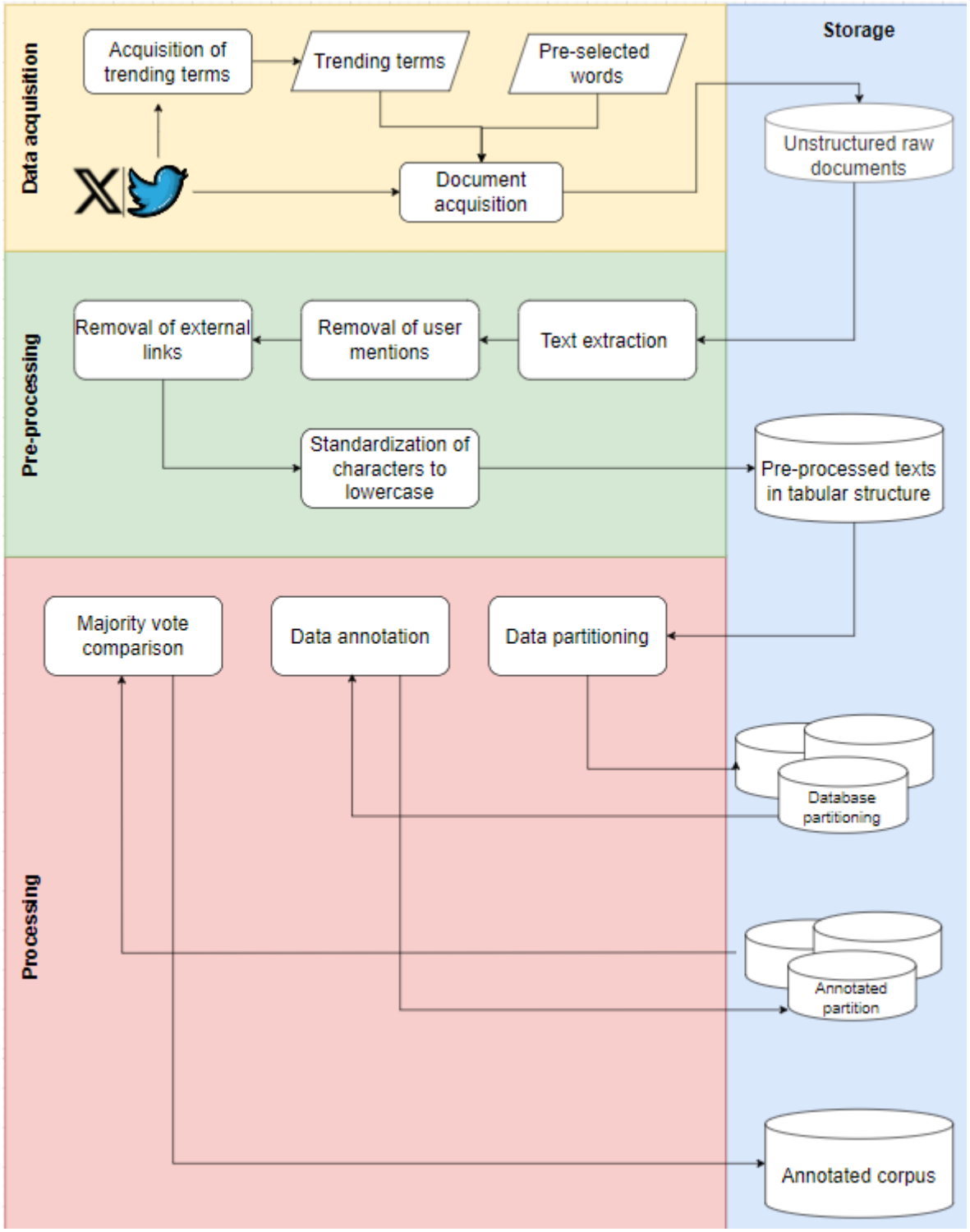

Figure 1. Methodology used in the creation of the TuPy dataset (adapted from Oliveira, 2024).

Table 1 illustrates the contribution of various sources to the TuPy-E dataset. Leite et al.'s work contributes the largest share, accounting for approximately 48.1% of the dataset, with around 11% of this being hate speech. Fortuna et al. contributes about 21.7% of the records, with a higher proportion of hate speech, roughly 22%. Meanwhile, TuPy and Vargas et al. make up approximately 22.9% and 16% of the dataset, respectively, with hate speech proportions of about 10.4% and 9.5%, respectively.

Table 1. Proportion of total documents and documents labeled as hate speech by data source used in the TuPy-E dataset.

Source	Total	Hate
Fortuna et al	5668	1228
Leite et al	21000	2319
TuPy	10000	1040
Vargas et al	7000	666
Total	43668	5253

3. Language Models

Based on the literature review conducted, we propose using BERT-based models, as this represents a key trend observed in the theoretical framework of this work. Additionally, tests with the GPT architecture were also conducted, given its relevance within the state-of-the-art language models.

The two versions of the BERTimbau model, BERTimbau Cased and GPortuguese-2, were refined using the TuPy-E dataset for two different classification tasks: binary classification for hate speech detection and categorical classification of hate speech.

BERTimbau is a pre-trained BERT model specifically designed for Brazilian Portuguese. It was pre-trained on an extensive Portuguese corpus, resulting in a robust representation of the language. We used BERTimbau in both its Base and Large configurations, each offering varying levels of adaptability and performance (Figure 23) (Souza, 2020) [11].

The Large variant, with its greater complexity, has 334 million parameters and can deliver superior performance across various tasks, although it requires more substantial computational resources. In contrast, BERTimbau-Base, with 109 million parameters, is suitable for a range of practical applications, including information extraction, document comprehension, and sentiment analysis in Brazilian Portuguese. Its lighter configuration makes it a viable option for scenarios with limited computational resources.

The GPortuguese-2 model, based on the GPT-2 Small architecture, was developed using Transfer Learning methodologies adapted for Portuguese (Guillou, 2020) [13].

Refinement of the aforementioned models was conducted using Google Colaboratory on a machine equipped with an NVIDIA A100 GPU, 1 CPU, 80 GB of RAM, and 40 GB of VRAM. We chose this online tool to provide an accessible computational environment, allowing other researchers to easily replicate this experiment.

The next section will provide a more detailed description of the experiments conducted.

3.1. Sampling and Splitting the Dataset for Training and Testing

To conduct the experiment effectively, we divided the dataset into distinct subsets for training, validation, and testing. The dataset used was TuPy-E, as described in Section 2, which includes a total of 43,668 annotated documents.

It is important to highlight that the dataset is imbalanced. It contains 5,252 documents classified as hate speech, with 9,367 instances of hate speech (considering multiple occurrences within the same document), compared to 34,301 documents that are not classified as hate speech.

This imbalance in class distribution can affect machine learning model performance, as many algorithms assume that training data is balanced by default (Fortuna, 2017). This assumption can introduce bias and impair the model's ability to generalize to new, unseen data.

To address this, we used the stratification feature from the Scikit-learn library to split the dataset into training and testing sets. Stratification ensures that the proportion of each class is maintained in both sets, which helps prevent bias in the evaluation of model performance.

We applied a standard 80/20 split, where 80% of the data was allocated for training the model, and the remaining 20% was reserved for testing and evaluating the model's performance.

3.2. Training and Testing Phases

For GPT-2, the training was conducted over three epochs with a batch size of 32. We used a loss function of 10^{-5} and the AdamW optimizer. Additionally, 500 warmup steps were implemented, and an L2 regularizer with a weight decay coefficient of 0.01 was applied.

For both BERT Base and BERT Large, training was performed over ten epochs, also with a batch size of 32. The loss function, optimizer, warmup steps, and L2 regularizer were the same as those used for GPT-2. These parameters were carefully chosen to enhance the performance of the models during training.

4. Results Analysis

This section presents the results of testing models on the TuPy-E dataset for two classification tasks: binary classification and categorical classification of hate speech.

4.1. Binary Classification of Hate Speech

We evaluated three refined language models—BERT-Base, BERT-Large, and GPT-2 Small—based on their respective versions: BERTimbau-Base, BERTimbau-Large, and GPortuguese-2. The goal was to assess their performance in binary hate speech classification for Portuguese text. The refinement process involved adapting these pre-trained models to the TuPy-E dataset.

As detailed in Table 2, both BERT-Base and BERT-Large achieved impressive performance in binary hate speech classification, with precision scores reaching 90%. This high precision indicates that these models are effective at identifying hate speech, balancing precision and recall well. This balance is crucial for reducing both false positives and false negatives, thus improving the overall reliability of the classification.

Table 1 – Performance metrics of different language models for binary hate speech classification

Model	Precision (Weighted)	Recall (Weighted)	F1 (Weighted)
BERT - Base	0.897	0.901	0.899
BERT - Large	0.901	0.907	0.903
GPT 2 - Small	0.888	0.892	0.890

Although GPT-2 Small had slightly lower performance metrics compared to the BERT models, including it in the evaluation added valuable insights into the proposed dataset. Despite a small decrease in performance, GPT-2 Small showed it is adaptable and effective in detecting hate speech, while being more resource-efficient.

The consistent F1 scores across all models highlight their reliability as a comprehensive evaluation metric. The similar performance of BERT-Base and BERT-Large indicates that BERT-Base is a practical choice, offering a less computationally demanding option without significantly compromising performance in hate speech detection.

4.2. Categorical Classification of Hate Speech

The results presented offer an analysis of performance metrics for the three models—BERT-Base, BERT-Large, and GPT-2 Small—applied to categorical hate speech classification. For each model, precision, recall, and

F1 scores are detailed for overall averages, as well as micro, macro, and weighted averages, along with values for each category. Table 3 provides the performance metrics for categorical classification across the BERT models.

Table 2 (A) - Performance metrics for BERT Base model in categorical hate speech classification.

Model	Category	Precision	Recall	F1	Suport
BERT Base	Aporophobia	1	0	0	16
	Capacitism	1	0	0	20
	Ageism	1	0	0	15
	Religious intolerance	0.25	0.11	0.15	19
	Lgbtphobia	0.85	0.67	0.75	171
	Misogyny	0.65	0.6	0.62	324
	Political	0.59	0.56	0.58	220
	Racism	0.29	0.27	0.28	62
	Body shame	0.58	0.54	0.56	54
	Xenophobia	0.41	0.31	0.35	78
	Others	0.56	0.49	0.52	909
	Not hate	0.92	0.93	0.92	7177
	Micro avg	**0.86**	**0.84**	**0.85**	**9065**
	Macro avg	**0.67**	**0.37**	**0.39**	
	Weighted avg	**0.85**	**0.84**	**0.84**	
	Samples avg	**0.86**	**0.85**	**0.85**	

Table 3 (B) - Performance metrics for BERT Large model in categorical hate speech classification.

Model	Category	Precision	Recall	F1	Suport
BERT Large	Aporophobia	0.75	0.19	0.30	16
	Capacitism	0.50	0.15	0.23	20
	Ageism	0.40	0.13	0.20	15
	Religious intolerance	0.27	0.16	0.20	19
	Lgbtphobia	0.78	0.75	0.76	171
	Misogyny	0.67	0.63	0.65	324
	Political	0.61	0.53	0.57	220
	Racism	0.39	0.42	0.40	62
	Body shame	0.78	0.65	0.71	54
	Xenophobia	0.39	0.22	0.28	78
	Others	0.62	0.46	0.53	909
	Not hate	0.91	0.94	0.93	7177
	Micro avg	**0.87**	**0.85**	**0.86**	**9065**
	Macro avg	**0.59**	**0.44**	**0.48**	
	Weighted avg	**0.85**	**0.85**	**0.85**	
	Samples avg	**0.87**	**0.86**	**0.86**	

BERT-Base had difficulty accurately identifying the categories of ageism, aporophobia, capacitism, and religious intolerance, resulting in lower precision, recall, and F1 scores for these categories. This is due to their lower representation in the dataset compared to more common categories. However, BERT-Base performed well in identifying categories like body shaming, political hate, LGBTphobia, misogyny, and xenophobia, maintaining a good balance between precision and recall. Racism was also a challenge for BERT-Base, with below-average performance due to its low support.

The average micro F1 score for BERT-Base is 0.85, reflecting overall strong performance with balanced precision and recall. However, the macro F1 score dropped to 0.39, indicating variability in performance across different categories, especially with lower recall values. The weighted average F1 score of 0.84 confirms the model's effectiveness in more supported categories.

BERT-Large shows improvements over BERT-Base, particularly in categories with high precision, recall, and F1 scores. It still struggles with ageism, aporophobia, capacitism, and religious intolerance, similar to BERT-Base. BERT-Large shows modest improvement in detecting racism but still faces challenges, as seen in its lower precision, recall, and F1 scores.

The average micro F1 score for BERT-Large increased slightly to 0.86, indicating better overall performance. The macro F1 score remains at 0.48, showing continued variability in performance across different categories. The weighted average F1 score is consistent at 0.85, demonstrating strong performance in more frequent categories.

Both BERT models effectively identify non-hate speech cases, as evidenced by high precision, recall, and F1 scores for the "non-hate" category, supported by significant overall support.

Table 4 displays the performance metrics for categorical classification of the GPT-2 Small model.

Table 4 - Performance metrics for the GPT-2 Small model in categorical hate speech classification.

Model	Category	Precision	Recall	F1	Suport
GPT 2 - *Small*	Aporophobia	1	0	0	16
	Capacitism	1	0	0	20
	Ageism	1	0	0	15
	Religious intolerance	1	0	0	19
	Lgbtphobia	0.76	0.67	0.71	171
	Misogyny	0.61	0.58	0.59	324
	Political	0.58	0.48	0.52	220
	Racism	1	0.05	0.09	62
	Body shame	0.71	0.59	0.64	54
	Xenophobia	0.5	0.19	0.27	78
	Others	0.65	0.36	0.47	909
	Not hate	0.91	0.92	0.92	7177
	Micro avg	**0.87**	**0.82**	**0.84**	**9065**
	Macro avg	**0.81**	**0.32**	**0.35**	
	Weighted avg	**0.86**	**0.82**	**0.83**	
	Samples avg	**0.89**	**0.83**	**0.83**	

The detailed analysis of GPT-2 Small reveals uneven performance across different hate speech categories. The model achieved 100% precision for ageism and aporophobia but failed to achieve recall for these categories. This suggests that while the model correctly identified the few relevant cases, it missed other existing instances.

For categories such as body shaming, LGBTphobia, political hate, misogyny, and others, the model performed reasonably well, with F1 scores ranging from 0.52 to 0.71. However, it struggled with racism, religious intolerance, and xenophobia, showing very low F1 scores. These results reflect significant difficulties in identifying these types of hate speech, compounded by low support values.

The model achieved very high precision for the "non-hate" category but had a proportionally lower recall, suggesting a tendency to misclassify some hate speech cases as non-hate.

With a micro F1 score of 0.85, GPT-2 Small showed a balanced performance between precision and recall for hate speech detection in Portuguese. However, the macro F1 score dropped to 0.35, indicating variability in performance across different classes, likely due to category imbalance. The weighted average F1 score of 0.83 underscores the model's effectiveness across various classes, despite a slight decrease compared to the BERT models.

When compared to results from other studies on the same task, the models developed in this work demonstrate performance comparable to the state-of-the-art in automated hate speech detection. Specifically, in the context of Portuguese, these models are among the best available to date.

It is important to note that the models developed in this research are fully compatible with variable importance tools like SHAP, making them potential instruments for analyzing hate speech characteristics. Although this study did not explore this area in depth due to technological limitations, it highlights it as a potential direction for future research. Figure 2 illustrates how variable explanations work in hate speech classification.

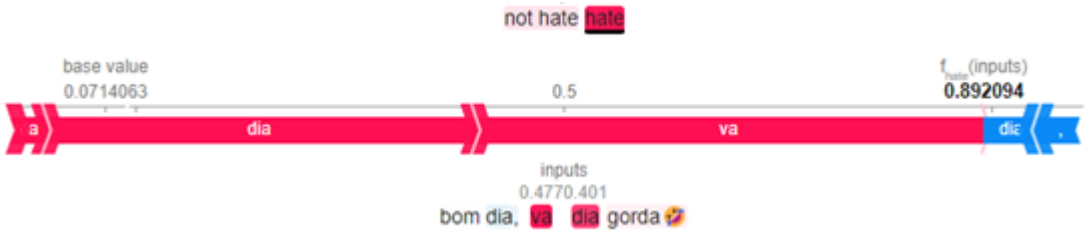

Figure 2. Methodology used in the creation of the TuPy dataset.

5. Accessibility

To ensure access to the models developed in this work and facilitate the reproduction of experiments, as well as to support future developments by other researchers—even those outside the AI field—an API has been created for online hate speech classification. Figure 3 illustrates how this application functions.

Figure 3. Hate speech classification API (adapted from Oliveira, 2024).

This study supports open-source policies by making all major products developed available online. The code can be found on GitHub, while the databases, models, and the classification API are accessible on Hugging Face.

References

[1] Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017, 512–515. https://doi.org/10.1609/icwsm.v11i1.14955
[2] Benesch, S. (2021). Dangerous Speech: A Practical Guide. Dangerous Speech Project.
[3] UNESCO. (2021). Addressing hate speech on social media: contemporary challenges. In United Nations Office on Genocide Prevention and the Responsibility to Protect (1st ed., Vol. 1, pp. 0–10). UNESDOC Digital Library..
[4] Wu, X. K., Zhao, T. F., Lu, L., & Chen, W. N. (2022). Predicting the Hate: A GSTM Model based on COVID-19 Hate Speech Datasets. Information Processing and Management, 59(4). https://doi.org/10.1016/j.ipm.2022.102998.
[5] Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). COUNTERING ONLINE HATE SPEECH. United Nations Educational, Scientific and Cultural Organization.
[6] Jahan, M. S., & Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. In Neurocomputing (Vol. 546). Elsevier B.V. https://doi.org/10.1016/j.neucom.2023.126232
[7] Fortuna, P. (2017). Automatic detection of hate speech in text: an overview of the topic and dataset annotation with hierarchical classes.
[8] Fortuna, P., Rocha Da Silva, J., Soler-Company, J., Wanner, L., & Nunes, S. (2019). A Hierarchically-Labeled Portuguese Hate Speech Dataset. https://github.com/t-davidson/hate-s
[9] Leite, J. A., Silva, D. F., Bontcheva, K., & Scarton, C. (2020). Toxic Language Detection in Social Media for Brazilian Portuguese: New Dataset and Multilingual Analysis. http://arxiv.org/abs/2010.04543
[10] Vargas, F., Carvalho, I., Góes, F., Pardo, T. A. S., & Benevenuto, F. (2022). HateBR: A Large Expert Annotated Corpus of Brazilian Instagram Comments for Offensive Language and Hate Speech Detection. https://www.statista.com/
[11] Souza, F. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese.
[12] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. http://arxiv.org/abs/1409.3215
[13] Guillou, P. (2020). GPorTuguese-2 (Portuguese GPT-2 small): a Language Model for Portuguese text generation (and more NLP tasks...).
[14] Oliveira, F. R. de. (2024). DISCURSO DE ÓDIO EM REDES SOCIAIS: UMA ABORDAGEM AUTOMATIZADA PARA IDENTIFICAÇÃO EM CONTEÚDO EM PORTUGUÊS