The Impact of Data Augmentation on the Hate Speech Detection in Portuguese Language

Félix Leonel V. da Silva, Artur Cerri, Ulisses B. Corrêa, Larissa A. de Freitas

CDTEC, Universidade Federal de Pelotas
96010-610 - Brazil
{flvdsilva, amcerri, ulisses, larissa}@inf.ufpel.edu.br

Abstract

Online communities allow users to establish a web presence, manage their identities, and stay connected with others. The internet has facilitated global outreach with just a click on the World Wide Web. However, the current landscape of online social media platforms are marred by various issues, with hate speech prominently taking center stage. Hate speech is characterized by hostile and malicious language driven by prejudice, targeting individuals or groups based on their innate, natural, or perceived characteristics. Detecting such speech is crucial for maintaining a safe online environment. This study examines the impact of dataset regularization techniques on the performance of BERTimbau-based models when applied to four Portuguese hate speech datasets: Fortuna et al. (2019), OFFCOMBR-2, ToLD-BR, and Hate-BR. Four Data Augmentation techniques are evaluated: Oversampling, Undersampling, Text Augmentation, and Synonym Replacement. Our experiments revealed that, apart from the Fortuna et al. (2019) dataset, the Data Augmentation techniques did not significantly enhance the performance of hate speech detection tasks.

Introduction

The history of online or digital social media is evidently a recent development, indicating the increasing use of computers to connect people. Social networks, such as Facebook, are dedicated to building and reflecting relationships within communities with similar interests. These platforms have evolved into mainstream environments for both teens and adults to exchange information, including private messages, share pictures, videos, and more. This transformation has opened up new avenues for personal profiling, enabling self-expression and the sharing of interests (Bandgar, 2014).

According to Mathew et al. (2018), the prevalence of online hate speech has contributed to horrific real-world hate crimes, such as the mass genocide of Rohingya Muslims, communal violence in Colombo, and the massacre in the Pittsburgh synagogue. Consequently, it is imperative to understand the diffusion of such hateful content in an online setting.

Detecting hate speech is a challenging task. The divergence in understanding regarding human labeling of hate speech underscores the increased difficulty of this classification for computer models. Various techniques, including Machine Learning (e.g., Logistic Regression and Decision Trees), Deep Learning (e.g., Convolutional Neural Network and Transformers), or combinations of these methods, can be employed for hate speech detection (Firmino, 2022).

This article is structured into the following sections: Theoretical Background, covering important concepts such as Hate Speech, Social Networks, Data Augmentation techniques, and Deep Learning; Related Works, explaining the references used in this work; Methodology, describing the approach used for the experiments, including datasets and Data Augmentation techniques; Analysis of the Results, presenting the outcomes of the experiments; and Final Remarks, summarizing key aspects of the work and providing a brief discussion about future works.

Theoretical Background

Natural Language Processing (NLP) is a field within Artificial Intelligence and Linguistics that focuses on enabling computers to understand expressions or words in human languages, known as natural languages. This field is subdivided into natural language understanding and natural language generation (Khurana et al., 2022), with Hate Speech Detection being a task falling under natural language understanding.

Hate Speech is defined as hostile and malicious speech driven by prejudice, targeting individuals or groups based on innate, real, or perceived characteristics. It expresses discriminatory, intimidating, disapproving, antagonistic, and prejudicial attitudes towards attributes such as gender, race, religion, ethnicity, color, nationality, disability, or sexual orientation. Hate speech aims to harm, dehumanize, harass, intimidate, demean, degrade, and victimize the targeted groups, fostering insensitivity and brutality against them (Cohen-Almagor, 2013). As mentioned earlier, social networks have become breeding grounds for the spread of hate speech. Therefore, this work utilizes datasets from the G1 platform (OFFCOMBR-2 (de Pelle and Moreira, 2017)), Twitter/X (Fortuna et al. (2019), ToLD-BR (Leite et al., 2020)), and Instagram (Hate-BR (Vargas et al., 2022)).

The G1 platform[1], created in 2006, serves as Globo's news website and was the initial journalistic content initiative designed for the digital realm. It aggregates journalism content from various Globo Group companies, presenting reports in text, photo, audio, and video formats. Twitter/X[2] is a communication service where friends, family, and colleagues can exchange quick and frequent messages containing photos, videos, links, and text. Instagram[3], established in 2010 and acquired by Facebook in 2012, is an online social network primarily focused on sharing photos, videos, and stories, featuring an efficient search engine.

Data Augmentation is a crucial strategy in Machine Learning and holds particular importance in NLP tasks. The Hate Speech Detection task faces challenges due to a scarcity of labeled datasets in Portuguese. Data Augmentation techniques are vital in increasing the available training data, improving model robustness and generalization capacity, and minimizing overfitting, especially when training data is limited (Pellicer, Ferreira, and Costa, 2023).

This work evaluates four Data Augmentation techniques: Oversampling, Undersampling, Text Augmentation, and Synonym Replacement. Oversampling involves increasing the number of instances or samples of the minority class until it matches the majority class by generating new instances or repeating some instances. Borderline-SMOTE is an example of an Oversampling method. Undersampling is the process of reducing the number of instances or samples of the majority class until it matches the number of the minority class, with common methods including tomek links, cluster centroids, and others (Mohammed, Rawashdeh, and Abdullah, 2020). Text Augmentation combines synonym replacement, random word swap, random character swap, and noise addition. Synonym Replacement selects n random words (excluding stopwords) and swaps them with randomly chosen synonyms. Random word swap selects two words randomly in the sentence and swaps their positions; this process is repeated n times (Wei and Zou, 2019).

Deep Learning (DL) enables computer models with multiple processing layers to learn data representations at various levels of abstraction. These methods have significantly advanced speech recognition, visual object recognition, and object detection. DL algorithms use backpropagation to indicate how a machine should adjust its internal parameters to calculate the representation in each layer based on the representation in the previous layer (LeCun, Bengio, and Hinton, 2015).

A groundbreaking development in machine learning architecture is the introduction of Transformers. This model architecture avoids recursion and relies entirely on an attention mechanism to establish global dependencies between input and output (Vaswani et al., 2017).

One notable application of the Transformer architecture is Bidirectional Encoder Representations from Transformers (BERT), designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning the left and right context in all layers. Consequently, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for various tasks, such as question-answering systems and linguistic inference, without substantially modifying the task-specific architecture (Devlin et al., 2019).

In the literature, several pre-trained BERT models for the Portuguese language exist, including BERTimbau (a BERT model from NeuralmindAI available in Base and Large versions (Souza, Nogueira, and Lotufo, 2020)), BioBERTpt (a BERT model pre-trained on medical and biomedical texts in Portuguese (Schneider et al., 2020)), BERTaú (pre-trained on search data from Itaú bank's virtual assistant (Finardi et al., 2021)), Portuguese BERT base cased Question Answering (trained on SQUAD v1.1 in Portuguese (Guillou, 2021)), BERTabaporu (designed for Brazilian Portuguese language and specifically tailored to Twitter data (da Costa et al., 2023)), Albertina (an enhanced iteration of the BERT model with a focus on the Portuguese language (Rodrigues et al., 2023)) and Sabiá-7B is a Portuguese language model developed by Maritaca Ai; the model accepts Only text inputs and generates text only. Sabiá7-b is an auto-regressive language model that uses the same architecture of LLaMA-1-7B; the model was trained on 7 billion tokens from the Portuguese subset of ClueWeb22 (Pires et al. (2023)).

Related Works

Among the works exploring Hate Speech Detection for the Portuguese language, with varied focuses, noteworthy contributions include de Pelle and Moreira (2017), Fortuna et al. (2019), Leite et al. (2020), Silva and Roman (2020), Silva and Freitas (2022), Scalercio and Freitas (2023) and Amadeus and Branco (2023), which introduces data augmentation for Portuguese.

In de Pelle and Moreira (2017), two datasets, OFFCOMBR-2 and OFFCOMBR-3, were created. The comments in these datasets were converted to lowercase and tested both in their original form and using n-grams (unigrams; unigrams and bigrams; unigrams, bigrams, and trigrams) as features. Support Vector Machines (SVM) and Na"ive Bayes (NB) served as classifiers with 10-fold cross-validation. Results averaged an F-measure of 0.80 for SVM and 0.75 for NB across the two datasets.

Fortuna et al. (2019) crafted a dataset with Portuguese tweets, splitting it into 90% for training and 10% for testing. Long Short-Term Memory (LSTM) was employed as a classifier with lowercase text, removal of punctuation and stopwords, and both cross and holdout validation. The result was a micro-average F-measure of 0.78 for LSTM.

The dataset ToLD-BR was introduced by Leite et al. (2020), divided into 80% for training, 10% for validation, and 10% for testing. Bag-of-Words (BoW) represented examples, and an AutoML model (BoW + AutoML) was built using auto-sklearn. For the BERT model, two versions of simple transformers were utilized: BERTimbau and mBERT. Results showed an F-measure of 0.74 for BoW + AutoML, 0.76 for BERTimbau, and 0.75 for m-BERT.

Silva and Roman (2020) utilized datasets from Fortuna et al. (2019), splitting them into 90% for training and 10%

[1]https://g1.globo.com

[2]https://twitter.com/

[3]https://www.instagram.com/

for testing. Classifiers NB, Logistic Regression (LR), SVM, and MultiLayer Perceptron (MLP) were applied with 10-fold cross-validation. The best result was an F-measure of 0.72 for BoW + SVM.

In Silva and Freitas (2022), datasets from Fortuna et al. (2019) and de Pelle and Moreira (2017) were used, split into 80% for training, 10% for validation, and 10% for testing. BERTimbau served as the model, and Data Augmentation techniques were applied to balance the datasets. The best result achieved was an F-measure of 0.91 with Oversampling.

Amadeus and Branco (2023) categorized Data Augmentation into three groups: Easy Data Augmentation (EDA), synonyms, and back translation. Three Portuguese datasets — TweetSentBR (Brum and das Graças Volpe Nunes, 2017), B2W open product reviews (Real, Oshiro, and Mafra, 2019), and Mercado libre data challenge 2019[4] — were utilized. SVM with specific parameters and fastText Portuguese Word Embedding model were employed. Synonym augmentation exhibited the best overall performance, achieving optimal experiment results.

Scalercio and Freitas (2023) used the Data Augmentation strategy, which consists of relocating and modifying adverbial structures whose position in the sentence is flexible. To do this, two structures with an adverbial function were chosen. They selected sentences in which an adverbial subordinate clause appears before the respective verb modified by it and sentences in which an adverbial syntagma appears before the verb is modified by it. In both transformations, a paraphrase is generated using the exact words as the source sentence but moving the target linguistic structure to after the subject, verb, and objects. The texts were annotated according to UD grammar, using the UDPipe syntactic annotator, which provides dependency trees that allow the two structures to be found. The UD syntactic labels OBL and ADVCL contain the cores of adverbial phrases and adverbial subordinate clauses. Once found, the word corresponding to the structure's core, together with all its children in the dependency tree, is moved to the front of the sentence after the terms that make up the subject, verb, and objects. They used three books from the Literature for all collections to evaluate the method: "Léo o pardo", "Madalena", and "Tubarão com faca nas costas". For each structure, the accuracy of the UDPipe annotator was calculated. Sixty random samples were chosen, 20 from each book; for the ADVCL label, the accuracy was 90%, and for the OBL label, 83.3%.

Methodology

The experiments carried out in this work used four datasets and BERTimbau Base (Souza, Nogueira, and Lotufo, 2020). The datasets were divided into 80% for training, 10% for validation, and 10% for testing. It is important to note that we used as hyperparameters 8 for the size of the batch, 4 epochs for training, CrossEntropy for the loss function, and AdamW as the optimizer.

[4]https://metatext.io/datasets/mercadolibre-data-challenge-2019

Datasets

For this work, the *datasets* were used: OFFCOMBR-2 (de Pelle and Moreira, 2017), Fortuna et al. (2019), ToLD-BR (Leite et al., 2020) and Hate-BR (Vargas et al., 2022). One of the reasons for choosing these four datasets was that one is from comments on a news site (OFFCOMBR-2), one is from comments on Instagram Hate-BR, and two are from tweets (Fortuna et al. (2019) and ToLD-BR), thereby obtaining a greater variety of text sizes and types for the experiments.

de Pelle and Moreira (2017) created the OFFCOMBR-2 with collected comments from the news site G1 in the categories of sport and news. The authors selected 1,250 comments, 419 comments with hate speech, and 831 comments without hate speech. They used the majority vote to decide whether a comment had hate speech.

The dataset of Fortuna et al. (2019) was collected from the social network Twitter/X between January and March 2017. The authors used Twitter/X's profile search API for keywords and hashtags such as sapatão or #LugarDeMulherENaCozinha. 29 specific profiles, 19 keywords, and 10 hashtags were observed. At the end of the process, 42930 tweets were collected. Two hundred tweets were chosen for each search instance, resulting in 5668 of the 42930 tweets. Each of the 5668 tweets was annotated by three annotators, with 1786 tweets with hate speech and 3882 tweets without hate speech, which classified each tweet as offensive or not.

The ToLD-BR was created by Leite et al. (2020). It contains tweets collected between July and August 2019. The authors used two strategies to collect the tweets: the first strategy was to search for keywords such as gay, little woman, and northeasterner, and the second strategy was to collect tweets that mentioned influential people such as former president Jair Bolsonaro and player Neymar, this method had no restrictions on keywords or hashtags, from all the tweets collected 21000 tweets were randomly selected. For the annotation process, 42 annotators classified 1500 comments each, and each tweet was classified by three annotators by majority vote, and they were classified as LGBTQ+phobia, obscene, insult, racism, misogyny, or xenophobia. In the end, 9245 tweets with hate speech and 11693 tweets without hate speech were obtained.

The Hate-BR was created by Vargas et al. (2022). It contains 7000 comments collected from Instagram. The comments were collected from six Instagram accounts of Brazilian politicians: three from the liberal party and three from the conservative party, four women and two men. The most popular posts from each account during the second half of 2019 were selected, with five posts for each account and 500 comments for each post. Only experts with high levels of education were selected to carry out the annotation to minimize bias and its negative impact on the results. In the end, 7000 comments were annotated, 702 with hate speech, and 6298 without hate speech.

Data Augmentation

There are several Data Augmentation libraries and techniques, such as TextAttack (Morris et al., 2020), NLPAug (Ma, 2019), and Easy Data Augmentation (EDA) (Wei and

Zou, 2019), which offer significant support for the process. However, these are general-purpose tools, and although they are highly valuable, they can face challenges when applied to languages with distinct and complex characteristics, such as Brazilian Portuguese.

The Synonym Replacement technique, commonly employed in Data Augmentation, is highly language-specific. In the case of Portuguese, for example, an appropriate synonym depends very much on the context in which the word is inserted, given the wealth of meanings and nuances in this language. Furthermore, the availability of an extensive and comprehensive library of synonyms, updated and adapted to the particularities of Portuguese, is still a challenge.

Inspired by the EDA technique devised by Wei and Zou (2019), we developed our own Data Augmentation library to Portuguese, implemented in the Python language. Due to its widespread use and proven effectiveness in improving performance in text classification tasks, the EDA technique proved to be an appropriate starting point for the design of our tool.

The EDA technique consists of four fundamental operations for data expansion: synonym replacement, random insertion, random exchange, and random deletion. For our library, we have adjusted and adapted these operations, as detailed below; an example of the application of the functions, original text: "nao esqueça que ele quer receber refugiados de guerra, vai islamizar o brasil" ["don't forget that he wants to receive war refugees, he's going to Islamize Brazil "].

1. Synonym replacement: We developed a function responsible for replacing words in the text with their respective synonyms in Brazilian Portuguese. To do this, we used a dictionary that compiles information from reliable sources, described in more detail in the following subsection, for example: "nao desmemorie que ele quer receber refugiados de combate, vai islamizar o brasil".
2. Random character swap: Unlike the original method, we implemented a function that performs random character swaps within words. This adjustment simulates typing errors, a common phenomenon in everyday texts, for example "nao eesqçua que ele quer recbeer refugiados de guerra, vai islamizar o brasil".
3. Random word swap: We created a function that randomly shuffles the words in a sentence to generate new variants of the same text while keeping the overall context intact, for example: "vai esqueça que ele quer receber refugiados de guerra, islamizar brasil o nao".
4. Noise addition: We developed a function to insert "noise" into texts by randomly inserting or removing characters. This practice creates a more challenging training scenario for NLP models, making them more robust, for example: "nao esquea que ee quer receber refugiados de guerra, vai islamizar o brasil".
5. Text Augmentation: We developed a function that uses synonym replacement, random character swap, random word swap, and noise addition, for example: "nao de que ele refugiados reecber quer deslembre guerra, vai islamizar o brasil".

The strategy adopted to build a set of synonyms was the process of web scraping on the website of the Common Orthographic Vocabulary of the Portuguese Language[5], combining the data extracted with the list of Portuguese language words made available by the Institute of Mathematics and Statistics of the University of São Paulo (IME-USP)[6]. We then used a second web scraping process on the Dicio[7] Online Portuguese Dictionary website, consolidating a rich set of synonyms. Even so, this approach has limitations, as it does not include checking the meaning and context of words, an aspect we plan to address in future library updates. The code for data augmentation techniques is publicly available at the following address: `https://github.com/amcerri/data-augmentation-GASPLN`.

Analysis of the Results

The following metrics were used to carry out the experiments: Accuracy, Balanced Accuracy, Recall, Precision, and F-Measure. The experiments used five configurations: Original, Oversampling, Undersampling, Synonym Replacement, and Text Augmentation, and four datasets OFFCOMBR-2 (de Pelle and Moreira, 2017), Fortuna et al. (2019), ToLD-BR (Leite et al., 2020) and Hate-BR (Vargas et al., 2022).

The Table 1 shows the results obtained from the experiments performed with the four datasets; for the Hate-BR, the results ranged from 0.97 to 0.86; for the Fortuna et al. (2019), the results ranged from 0.94 to 0.85; for the OFFCOMBR-2 the results ranged from 0.93 to 0.89 and for the ToLD-BR the results ranged from 0.91 to 0.88. Meanwhile, Table 2 shows some examples that were misclassified in the model, mainly by the Undersampling and Text Augmentation configurations, which were the two configurations that achieved the worst results overall among the four datasets, especially Undersampling, which performed considerably less well than the other configurations.

The OFFCOMBR-2 (de Pelle and Moreira, 2017) dataset showed remarkable results, with superior performance in the Original, Oversampling, and Synonym Replacement configurations. At the same time, it obtained less satisfactory results in the UnderSampling and Text Augmentation configurations. Further analysis reveals that the positive results in the configurations mentioned above can be attributed to the preservation of the original dataset, the increased representation of the minority class through Oversampling, and the improved generalization through synonym replacement. In contrast, the unfavorable results in the UnderSampling configurations indicate that excluding examples from the majority class compromised the model's ability to capture crucial nuances, leading to inferior performance. In addition, the Text Augmentation technique, by randomly introducing character and word swaps, resulted in texts that possibly became challenging for the model to classify, highlighting OFFCOMBR-2's sensitivity to semantic changes not aligned with the original characteristics of the data.

[5]https://voc.cplp.org/

[6]https://www.ime.usp.br/ pf/dicios/

[7]https://www.dicio.com.br/

Table 1: Results obtained with the experiments.

Configurations	Dataset	Accuracy	Balanced Accuracy	Recall	Precision	F-Measure
Original	OFFCOMBR-2	**0.93**	0.92	**0.93**	**0.93**	**0.93**
Original	Fortuna et al. (2019)	0.93	0.91	0.93	0.93	0.93
Original	ToLD-BR	**0.91**	**0.91**	**0.91**	**0.91**	**0.91**
Original	Hate-BR	0.97	0.90	0.97	0.97	0.97
OverSampling	OFFCOMBR-2	**0.93**	0.92	**0.93**	**0.93**	**0.93**
OverSampling	Fortuna et al. (2019)	**0.94**	**0.94**	**0.94**	**0.94**	**0.94**
OverSampling	ToLD-BR	0.91	0.90	0.91	0.91	0.91
OverSampling	Hate-BR	**0.97**	**0.93**	**0.97**	**0.97**	**0.97**
UnderSampling	OFFCOMBR-2	**0.89**	**0.89**	**0.89**	0.90	0.90
UnderSampling	Fortuna et al. (2019)	**0.85**	**0.87**	**0.85**	**0.87**	**0.85**
UnderSampling	ToLD-BR	**0.88**	**0.89**	**0.88**	**0.89**	**0.88**
UnderSampling	Hate-BR	**0.86**	**0.87**	**0.86**	**0.92**	**0.88**
Text Augmentation	OFFCOMBR-2	**0.89**	**0.88**	**0.89**	**0.89**	**0.89**
Text Augmentation	Fortuna et al. (2019)	0.92	0.90	0.92	0.92	0.92
TextAugmentation	ToLD-BR	0.89	0.89	0.89	0.89	0.89
Text Augmentation	Hate-BR	0.95	0.87	0.95	0.95	0.95
Synonym Replacement	OFFCOMBR-2	**0.93**	**0.93**	**0.93**	**0.93**	**0.93**
Synonym Replacement	Fortuna et al. (2019)	0.92	0.91	0.92	0.92	0.92
Synonym Replacement	ToLD-BR	**0.91**	**0.91**	**0.91**	**0.91**	**0.91**
Synonym Replacement	Hate-BR	0.96	0.90	0.96	0.96	0.96

In the context of Fortuna et al. (2019) dataset, a nuanced examination reveals that class imbalance has a pronounced impact on this dataset compared to others. Despite a marginal difference, the OverSampling configuration consistently outperformed alternative configurations across all metrics. Conversely, the UnderSampling configuration exhibited inferior results in comparison to its counterparts. The rationale behind the disparity lies in the fact that UnderSampling, as elucidated earlier, entails the removal of instances from the dataset to rectify class imbalances. While aiming for balance, this process inadvertently diminishes the pool of examples available for classification, potentially discarding pivotal texts crucial for diverse and comprehensive classification. The consequential loss of textual diversity may serve as a plausible explanation for the suboptimal performance associated with the UnderSampling configuration. These findings underscore the sensitivity of the dataset to class distribution nuances and underscore the critical role of effective sampling strategies in enhancing classification outcomes.

A thorough analysis of the ToLD-BR dataset (Leite et al., 2020) suggests that the experimental outcomes exhibited minimal variability across different configurations. Intriguingly, the class imbalance in this dataset did not exert a discernible impact on the results. Notably, the Original and Synonym Replacement configurations yielded the most favorable outcomes. This consistent performance, irrespective of the applied configurations, prompts an exploration of the dataset's inherent characteristics. The resilience of the ToLD-BR dataset to class imbalances implies that the model's performance remains robust despite skewed class distributions. The success of the Original configuration suggests that the dataset's natural distribution provides sufficient information for effective learning. Additionally, the efficacy of the Synonym Replacement configuration implies that subtle semantic manipulations contribute positively to the model's generalization.

In the context of the Hate-BR dataset (Vargas et al., 2022), a nuanced examination following the experiments reveals that optimal results were achieved with the OverSampling configuration. Notably, the marginal performance difference between configurations suggests that class imbalance has a limited impact on the model's classification performance for this dataset. Parallel to observations in other datasets, the UnderSampling configuration consistently produced the least favorable results, emphasizing its inadequacy due to information loss. The dominance of the OverSampling configuration implies that replicating or generating additional instances of the minority class is beneficial for enhancing the model's ability to discern patterns within the dataset. The relatively minor performance distinction across configurations underscores the dataset's resilience to class imbalances, implying that the model can adeptly navigate classification tasks despite skewed class distributions. Conversely, the recurring poor performance associated with UnderSampling reinforces the idea that diminishing the dataset's information content adversely affects the model's learning capacity.

When comparing the results of the four datasets, it is essential to consider the difference between them in terms of size, type, and quantity of text. The results were better with the Hate-BR dataset (Vargas et al., 2022). The BERTimbau model (Souza, Nogueira, and Lotufo, 2020) performed better when dealing with the classification task in this dataset. The comments in the Hate-BR dataset should be better worded, and abbreviations should be avoided, as poor wording and the use of abbreviations were identified as one of the main factors contributing to a higher number of errors during the experiments. On the other hand, the configuration that resulted in the weakest performance was the one that used the UnderSampling technique on all four datasets. Still, this can be explained by the undersampling method, which removes

sentences from the majority class, which can lead to the loss of crucial information for the model's classification process.

Final Remarks

This study aimed to detect hate speech using BERTimbau (Souza, Nogueira, and Lotufo, 2020) as a classifier. Four Portuguese language datasets were employed in the experiments, revealing that the Hate-BR (Vargas et al., 2022) dataset outperformed the other three datasets across all metrics (Accuracy, Balanced Accuracy, Recall, Precision, and F-Measure). Notably, the application of Data Augmentation had minimal impact on the results, indicating that class imbalance did not significantly affect the outcomes. Despite the limited impact, the experiments yielded higher results than those reported in the related works with the four datasets.

In future works, we aspire to explore other NLP tasks, including Sentiment Analysis, Sarcasm and Irony Detection, and Fake News Detection, employing the Data Augmentation technique.

Additionally, we aim to leverage the capabilities of the LLM model Sabiá-7B for detecting hate speech, recognizing its efficient performance in handling the intricacies of language. This application expansion aligns with our commitment to harnessing advanced natural language processing models for diverse linguistic tasks. We also plan to incorporate back translation functionality in Data Augmentation, which is pending implementation at the moment. To achieve this, we are considering leveraging Transformer models for the necessary translations in the back translation process. Finally, we plan to replace the current dictionary with WordNet to preserve sentence semantics better.

Acknowledgments

This research was supported by the Institutional Scholarship Program for Initiation in Technological Development and Innovation - PIBITI/CNPq, in the project Aspect-Based Sentiment Analysis Using Deep Learning: a Proposal Applied to the Portuguese Language.

References

Amadeus, M., and Branco, P. 2023. Performance of data augmentation methods for brazilian portuguese text classification.

Bandgar, B. 2014. Role of socia network in recent era. *International Journal of Research in Ccomputer Science and Management Vol. 1(1), January 2014 (ISSN: 2321-8088)* 1:2321–8088.

Brum, H. B., and das Graças Volpe Nunes, M. 2017. Building a sentiment corpus of tweets in brazilian portuguese.

Cohen-Almagor, R. 2013. Freedom of expression v. social responsibility: Holocaust denial in canada. *Journal of Mass Media Ethics* 28:42–56.

da Costa, P. B.; Pavan, M. C.; dos Santos, W. R.; da Silva, S. C.; and Paraboni, I. 2023. BERTabaporu: assessing a genre-specific language model for Portuguese NLP. In *Recents Advances in Natural Language Processing (RANLP-2023)*.

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

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Finardi, P.; Viegas, J. D.; Ferreira, G. T.; Mansano, A. F.; and Caridá, V. F. 2021. Bertaú: Itaú bert for digital customer service. In *arXiv*.

Firmino, A. A. 2022. *Uma abordagem para detecção de discurso de ódio utilizando aprendizado de máquina baseado em cruzamento de idiomas.* Ph.D. Dissertation, Universidade Federal de Campina Grande.

Fortuna, P.; Rocha da Silva, J.; Soler-Company, J.; Wanner, L.; and Nunes, S. 2019. A hierarchically-labeled Portuguese hate speech dataset. In Roberts, S. T.; Tetreault, J.; Prabhakaran, V.; and Waseem, Z., eds., *Proceedings of the Third Workshop on Abusive Language Online*, 94–104. Florence, Italy: Association for Computational Linguistics.

Guillou, P. 2021. Portuguese bert base cased qa (question answering), finetuned on squad v1.1.

Khurana, D.; Koli, A.; Khatter, K.; and Singh, S. 2022. Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications* 82(3):3713–3744.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *nature* 521(7553):436.

Leite, J. A.; Silva, D. F.; Bontcheva, K.; and Scarton, C. 2020. Toxic language detection in social media for brazilian portuguese: New dataset and multilingual analysis. *CoRR* abs/2010.04543.

Ma, E. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

Mathew, B.; Dutt, R.; Goyal, P.; and Mukherjee, A. 2018. Spread of hate speech in online social media.

Mohammed, R.; Rawashdeh, J.; and Abdullah, M. 2020. Machine learning with oversampling and undersampling techniques: Overview study and experimental results. 243–248.

Morris, J. X.; Lifland, E.; Yoo, J. Y.; Grigsby, J.; Jin, D.; and Qi, Y. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp.

Pellicer, L. F. A. O.; Ferreira, T. M.; and Costa, A. H. R. 2023. Data augmentation techniques in natural language processing. *Applied Soft Computing* 132:109803.

Pires, R.; Abonizio, H.; Almeida, T. S.; and Nogueira, R. 2023. Sabiá: Portuguese large language models. In Naldi, M. C., and Bianchi, R. A. C., eds., *Intelligent Systems*, 226–240. Cham: Springer Nature Switzerland.

Table 2: Some examples that the model incorrectly classified in the four datasets.

Text	Dataset	Real Label	Predicted Label
"Sejam honestos aprovem o projeto original vamos acabar com esta farra no Brasil" "Be honest, approve the original bill and let's put an end to this spree in Brazil"	OFFCOMBR-2	0	1
"Estrategia VERMELGA ACEFALA terceirizar a culpa xingar e denegrirKKKK TCHAU QUERIDOS" "VERMELGA BRAINLESS strategy outsource the blame curse and denigrateKKKK BYE DEAR"	OFFCOMBR-2	1	0
"Magra - gorda , com ou sem estrias , feia - bonita ... não importa porque vocês são perfeitas! Feliz dia da mulher" "Thin - fat, with or without stretch marks, ugly - beautiful ... it doesn't matter because you're perfect! Happy Women's Day"	Fortuna et al. (2019)	0	1
"já tava na hora de acabar com esse ciência sem fronteira um bando de pobre indo viajar pra europa no mesmo avião que eu?!????" "it's about time to put an end to this science without frontiers - a bunch of poor people going to Europe on the same plane as me?! ????"	Fortuna et al. (2019)	1	0
"Alertas do Inpe indicam alta de 40% em desmate na Amazônia' O que são esses alertas, quais são? O que eles realmente indicam? E os falsos positivos, qual a porcentagem? Que métrica é usada para definir um alerta? Algum cientista por aqui para responder?" "Inpe alerts indicate 40% increase in deforestation in the Amazon' What are these alerts, what are they? What do they really indicate? And what percentage of false positives? What metric is used to define an alert? Any scientists here to answer?"	ToLD-BR	0	1
"Lady murphy sempre atenta quando estou atrasada, ela faz questão de me mandar um uber bem burro, daqueles que até estanca o carro quando tá nervoser" "Lady murphy is always on the lookout when I'm late, she makes a point of sending me a really dumb uber, the kind that even stops the car when you're nervous"	ToLD-BR	1	0
"VOCÊ TINHA DE TER PENA É QUANDO VÁRIOS BRASILEIROS MORRERAM NOS CORREDORES DOS HOSPITAIS PÚBLICOS POR CAUSA DAS ROUBALHEIRAS DOS CURRUPTOS, MAS VOCÊ VELHA CANALHA TAMBÉM FAZIA PARTE DO ESQUEMA SEGUNDO Á OAS" "YOU HAD TO FEEL SORRY WHEN SEVERAL BRAZILIANS DIED IN PUBLIC HOSPITAL CORRIDORS BECAUSE OF THE THIEVERY OF THE CORRUPT, BUT YOU OLD SCOUNDREL WAS ALSO PART OF THE SCHEME ACCORDING TO OAS"	Hate-BR	0	1
"Eu tenho pena é de você não ter sido da época do grande General Pinochet que lamentavelmente deixou sobrar este resquício de desgraça" "My pity is that you weren't around during the time of the great General Pinochet, who unfortunately left this remnant of disgrace"	Hate-BR	1	0

Real, L.; Oshiro, M.; and Mafra, A. 2019. B2w-reviews01 - an open product reviews corpus.

Rodrigues, J.; Gomes, L.; Silva, J.; Branco, A.; Santos, R.; Cardoso, H. L.; and Osório, T. 2023. Advancing neural encoding of portuguese with transformer albertina pt-*.

Scalercio, A., and Freitas, C. 2023. Proposta e avaliação linguística de técnicas de aumento de dados. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, 207–223. Porto Alegre, RS, Brasil: SBC.

Schneider, E. T. R.; de Souza, J. V. A.; Knafou, J.; Oliveira, L. E. S.; Copara, J.; Gumiel, Y. B.; de Oliveira, L. F. A.; Paraiso, E. C.; Teodoro, D.; and Barra, C. M. C. M. 2020. BioBERTpt - a Portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*.

Silva, F. L. V., and Freitas, L. A. 2022. Brazilian portuguese hate speech classification using bertimbau. *The International FLAIRS Conference Proceedings* 35.

Silva, A., and Roman, N. 2020. Hate speech detection in portuguese with naïve bayes, svm, mlp and logistic regression. In *Proceedings of the 17th Encontro Nacional de Inteligência Artificial e Computacional*.

Souza, F.; Nogueira, R.; and Lotufo, R. 2020. Bertimbau: pretrained bert models for brazilian portuguese. In *Proceedings of the 9th Brazilian Conference on Intelligent Systems*.

Vargas, F.; Carvalho, I.; Rodrigues de Góes, F.; Pardo, T.; and Benevenuto, F. 2022. HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In Calzolari, N.; Béchet, F.; Blache, P.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 7174–7183. Marseille, France: European Language Resources Association.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need.

Wei, J., and Zou, K. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*

(EMNLP-IJCNLP), 6382–6388. Hong Kong, China: Association for Computational Linguistics.