An Approach for the Automatic Detection of Prejudice in Instant Messaging Applications

Melissa Sousa, Fernanda Nascimento, Gustavo Martins,
José Maria Monteiro and Javam Machado
Computer Science Department, Federal University of Ceará, Brazil

Keywords: Prejudice, Datasets, Automatic Detection.

Abstract: Instant messaging applications have revolutionized communication, making it more accessible and efficient. However, they have also facilitated the widespread dissemination of prejudiced media content. In this context, the rapid and effective detection of prejudice in texts shared via messaging apps is crucial for promoting a healthy, diverse, and tolerant communicative environment. Few prejudice detection methods have been specifically developed for instant messaging platforms. Moreover, the development of effective methods requires labeled datasets containing prejudiced messages disseminated on these platforms, as user expressions differ significantly from those on other social networks like Facebook, Instagram, and X. However, we have not found any datasets containing prejudiced messages extracted from WhatsApp or Telegram. This work presents two publicly available labeled datasets, named PrejudiceWhatsApp.Br and PrejudiceTelegram.Br, consisting of Brazilian Portuguese (PT-BR) messages collected from public groups on WhatsApp and Telegram, respectively. Additionally, we developed a dictionary of prejudiced words for Brazilian Portuguese, named PrejudicePT-br, comprising 842 words organized into nine categories. Finally, we built a dictionary-based machine learning model to automatically detect prejudice in WhatsApp and Telegram messages. We conducted a series of text classification experiments, combining two feature extraction methods, three distinct token generation strategies, two preprocessing approaches, and nine classification algorithms to classify texts into two categories: prejudiced and non-prejudiced. Our best results achieved an F1-score of 0.86 for both datasets, demonstrating the feasibility of the proposed approach.

1 INTRODUCTION

In recent years, the growing popularity of instant messaging applications has significantly transformed the way we produce, share, and consume information. In Brazil, WhatsApp stands out as one of the most widely used applications, with over 165 million users (de Sá et al., 2023b). Similarly, Telegram has also experienced remarkable growth in 2022, with the proportion of smartphones that have the application installed increasing from 45% to 60% in just one year (de Sá et al., 2023a). The widespread adoption of these applications can be attributed to their versatility and ease of use. Additionally, they offer a particularly relevant feature: public chat groups. These groups, accessible via invitation links, are generally organized around specific themes, such as politics, sports, finance, or education. Both WhatsApp and Telegram allow users to join hundreds of groups simultaneously, thereby connecting with thousands of other users in an integrated manner. However, while these applications promote fast and efficient communication, the lack of adequate control and regulation makes them conducive to the large-scale dissemination of prejudiced discourse.

The study of prejudice as an independent scientific concept began to gain attention from psychologists in the 1920s. Since then, numerous studies have systematically explored the causes and consequences of this phenomenon. One of the leading scholars in this field was the American psychologist Gordon Allport, who published his renowned book The Nature of Prejudice in 1954 (CROCHÍK, 1997). Allport emphasized the influence of personality traits, emotions, and cognitions on the development of prejudice. He defined prejudice as "a feeling, favorable or unfavorable, toward a person or thing prior to, or not based on, actual experience". However, much of contemporary re-

Sousa, M., Nascimento, F., Martins, G., Monteiro, J. M., Machado and J.
An Approach for the Automatic Detection of Prejudice in Instant Messaging Applications.
DOI: 10.5220/0013555100003967
In *Proceedings of the 14th International Conference on Data Science, Technology and Applications (DATA 2025)*, pages 108-119
ISBN: 978-989-758-758-0; ISSN: 2184-285X

search tends to agree that one of the most significant factors related to prejudice is its historical and social construction. In this regard, Aronson defined prejudice as "a hostile or negative attitude toward a particular group, based on distorted or incomplete generalizations" (GOLDSTEIN, 1983). Similarly, Mezan (MEZAN, 1998) described prejudice as "a set of beliefs, attitudes, and behaviors that attributes a negative characteristic to any member of a given human group, solely based on their belonging to that group; this characteristic is perceived as essential, defining the nature of the group, and therefore adheres indelibly to all individuals within it".

Prejudice is one of the most cruel forms of oppression and discrimination among individuals in contemporary society. It also functions as a form of self-punishment, as it fosters a sense of guilt in the victim. Any form of prejudice in human relationships is detrimental to the development of a just, diverse, and inclusive society. The violence of prejudice fosters isolation among individuals and subtly instills distrust among peers.

In this context, the rapid and effective detection of prejudiced discourse shared on instant messaging applications becomes fundamentally important for building a democratic and tolerant communicational environment. However, despite this scenario, there are few detection methods specifically developed for these platforms. Furthermore, for efficient methods to be developed, it is essential to have labeled datasets containing prejudiced messages that have been disseminated through these applications, as the way users express themselves differs significantly from public social networks such as Facebook, Instagram, and X (Rosenfeld et al., 2018). Nevertheless, we have not found any datasets containing prejudiced messages extracted from WhatsApp or Telegram.

Thus, to address this gap, we constructed two publicly available, anonymized, and labeled datasets consisting of messages in Brazilian Portuguese (PT-BR) collected from public WhatsApp and Telegram groups, respectively. These datasets, which contain prejudiced messages, were named PrejudiceWhatsApp.Br and PrejudiceTelegram.Br. Additionally, we developed a dictionary of prejudiced words for Brazilian Portuguese, named PrejudicePT-br, which comprises 842 words organized into nine categories.

Subsequently, based on the PrejudicePT-br dictionary, we propose an approach for the automatic detection of prejudiced messages. Finally, a series of text classification experiments was conducted, combining two different feature extraction methods (Bag of Words (BoW) and TF-IDF (Term Frequency – Inverse Document Frequency)), three different tokenization strategies (unigrams, bigrams, and trigrams), two preprocessing approaches (no preprocessing and removal of stop words with lemmatization), and nine classification algorithms to categorize texts into two classes: prejudiced texts (associated with some form of prejudice) and non-prejudiced texts. These experiments were performed using both PrejudiceWhatsApp.Br and PrejudiceTelegram.Br.

The results indicate that it is possible to efficiently identify prejudiced messages. The best models achieved an F1-score of 0.86 for both the PrejudiceTelegram.Br and PrejudiceWhatsApp.Br datasets, demonstrating the feasibility of the proposed approach. To the best of our knowledge, no previous study has publicly released labeled datasets containing prejudiced messages from WhatsApp and Telegram, nor has any research systematically evaluated strategies for the automatic detection of prejudice in this context.

The remainder of this article is organized as follows. Section 2 discusses the main related works. The PrejudiceWhatsApp.Br and PrejudiceTelegram.Br datasets are presented in Section 3. Section 4 details the experiments conducted to evaluate different predictive models for detecting prejudice in instant messaging applications. Section 5 discusses the obtained results. Finally, Section 6 presents the conclusions drawn from this study and outlines possibilities for future research.

2 RELATED WORKS

Initially, it is essential to differentiate between prejudiced, pejorative, offensive, toxic language, and hate speech. Prejudiced language involves a negative stance toward a particular group, based on distorted or incomplete generalizations. For example: "Women use the right side of the brain more, which is why they are more intuitive, emotional, multitasking, and have lower skills in exact sciences". Pejorative language employs terms that devalue something or someone, such as: "That is so amateurish" or "What a crappy chair". Offensive language includes expressions that directly insult or cause discomfort, such as: "You are an idiot". Toxic language fosters a communication style that creates a negative environment, with or without offensive words. For example: "No one can stand you". Hate speech consists of expressions that promote discrimination or violence, often with more severe consequences. For instance: "This group should be banned". Naturally, these concepts may sometimes overlap. For example, a toxic message may include offensive expressions.

In (Dinu et al., 2021), the authors investigated the following classification task: given a word and a tweet in which the word appears, the goal was to determine whether the word was used pejoratively in that tweet. The experiments were conducted using two datasets (PEJOR1 and PEJOR2), which contain 944 and 313 pairs (tweet and word), respectively. The best model achieved an F1-score of 0.864. Additionally, the study introduced a multilingual lexicon of pejorative terms for English, Spanish, Italian, and Romanian, which was constructed based on online dictionaries.

HateBR, a corpus of comments containing hate speech and offensive language in Portuguese, collected from Instagram, was introduced in (Vargas et al., 2022). The corpus consists of 7,000 annotated comments, classified according to three different layers: binary classification (offensive vs. non-offensive comments), Offensiveness level classification (highly, moderately, and mildly offensive comments), and nine categories of hate speech (including xenophobia, racism, homophobia, sexism, religious intolerance, partisanship, advocacy for dictatorship, antisemitism, and fatphobia). Additionally, a series of experiments was conducted using HateBR to evaluate different machine learning algorithms for offensive language and hate speech classification. The best models achieved an F1-score of 0.85 for offensive language detection and 0.78 for hate speech detection.

A lexicon of offensive expressions in Portuguese, annotated with contextual information and named MOL (Multilingual Offensive Lexicon), was introduced in (Vargas et al., 2021). Additionally, based on the MOL lexicon, the authors proposed a novel approach for detecting offensive language and hate speech in social networks. The experiments were conducted using the HateBR dataset, and the best models achieved an F1-score of 0.88 for offensive language detection and 0.85 for hate speech detection.

In (Leite et al., 2020), the authors introduced a dataset named ToLD-Br (Toxic Language Dataset for Brazilian Portuguese), consisting of tweets in Portuguese annotated as toxic or non-toxic, as well as different types of toxicity. A total of 21,000 tweets was manually labeled into seven categories: non-toxic, LGBTQ+phobia, obscenity, insult, racism, misogyny, and xenophobia. Subsequently, the authors used ToLD-Br to evaluate the performance of BERT-based models for the automatic classification of toxic comments. The best model achieved a macro-F1 score of 0.76.

An affective dictionary for Brazilian Portuguese, named AffectPT-br, was proposed in (Carvalho et al., 2018). AffectPT-br was constructed based on the LIWC 2015 dictionary in English. Additionally, the authors evaluated the use of AffectPT-br and LIWC2007pt for polarity classification. The experiments were conducted on two distinct datasets: My Dear Diary (Meu Querido Diário in Portuguese - MQD) and TAS-PT. The best results achieved an F1-score of 0.715 for the MQD dataset and 0.960 for the TAS-PT dataset, both obtained using AffectPT-br.

The study presented in (Carvalho et al., 2023) utilized four publicly available datasets and seven machine learning algorithms to evaluate the use of the LIWC2015 dictionary compared to its previous version, LIWC2007, in polarity classification. The results demonstrated that LIWC2015 outperforms LIWC2007 in this task. Additionally, an emotional dictionary specifically designed for political texts in the German language was introduced and evaluated in (Widmann and Wich, 2023). The study compared the proposed dictionary with generic dictionaries by training different machine learning models. The results indicated that all customized approaches outperformed widely used pre-existing dictionaries in measuring emotional language in German political discourse.

The automatic classification of polarity in posts related to anxiety on a Chinese social media platform was investigated in (Zhu et al., 2024). Linguistic features of the posts, extracted using the Simplified Chinese–Linguistic Inquiry and Word Count (SC-LIWC) dictionary, were used to train a TextCNN-based model. The experimental results indicate that the proposed approach outperforms traditional methods in identifying the sentiment polarity of anxiety-related posts on Chinese social media. Similarly, in (Sert and Ülker, 2023), the authors examined how the combination of machine learning methods and LIWC can be used to detect mental disorders.

The study presented in (Taso et al., 2023) describes a series of experiments based on the Implicit Association Test (IAT) from psychology, which were used to identify and quantify biases in a Portuguese-language Word Embedding (WE). To achieve this, the authors utilized a GloVe model trained on a collection of Internet corpora. The results revealed that various common sense and gender stereotypes can be found in the WE, highlighting the importance of discussing the impact of language models on society.

In (Bahgat et al., 2022), the authors introduced LIWC-UD, an extension of the LIWC dictionary that incorporates terms from the Urban Dictionary. While the original LIWC contains 6,547 unique entries, LIWC-UD consists of 141,000 unique terms, which were automatically categorized into LIWC categories with high confidence using a BERT classifier.

Table 1: Main Characteristics of Related Works.

Work	Task	Public Dataset	Classifiers
[Dinu et al., 2021]	Detection of Pejorative Language	Yes	KNN, SVM e MLP
[Vargas et al., 2022]	Detection of Offensive Language and Hate Speech	Yes	NB, SVM, MLP e LR
[Vargas et al., 2021]	Detection of Offensive Language and Hate Speech	Yes	SVM, MNB, MLP e LSTM
[Leite et al., 2020]	Detection of Toxic Language	Yes	BERT
[CVarvalho et al., 2018]	Polarity Detection	Yes	SVM, MNB, RF e J48
[This Work]	Detection of Prejudiced Language	Yes	LR, BNB, MNB, KNN, LSVM, SGD, RF, GB e MLP

Table 1 presents the main characteristics of the related works, including the "Task" or problem investigated, whether the study provides publicly available datasets, and the classifiers evaluated in the experiments. It is important to note that, unlike previous studies, the present work focuses on a specific task, namely the detection of prejudiced language. Moreover, it distinguishes itself by utilizing datasets collected from WhatsApp and Telegram, which have not been explored in prior research.

3 THE PROPOSED DATASETS

To develop an automatic detector for prejudiced texts in the context of instant messaging applications, it is essential to utilize a large-scale labeled dataset composed of messages in Brazilian Portuguese (PT-BR) that have circulated on these platforms. However, no existing corpus with these characteristics has been identified. To bridge this gap, we constructed two datasets, named PrejudiceWhatsApp.Br and PrejudiceTelegram.Br, consisting of messages collected from public groups on WhatsApp and Telegram, respectively. For this purpose, we followed the guidelines proposed by (Rubin et al., 2015) for the construction of a corpus designed for classification tasks.

3.1 Data Collection

The collection of messages from WhatsApp and Telegram was conducted using the BATMAN platform[1] between August 1, 2022, and December 31, 2022 (de Sá et al., 2023b). On WhatsApp, a total of 813,106 unique messages (i.e., non-repetitive) were collected from 179 public groups. On Telegram, 767,847 unique messages were captured from 150 public groups and/or channels. Manually labeling such a large volume of messages is unfeasible. Therefore, a strategy to reduce the number of messages to be annotated is necessary. To address this, we adopted an approach based on the use of a word dictionary.

3.2 The PrejudicePT-br Dictionary

We developed a dictionary of prejudiced words for Brazilian Portuguese, named PrejudicePT-br, which was constructed using the LIWC2015pt, AffectPT-br, MOL, and Wiktionary dictionaries as sources. PrejudicePT-br consists of 842 words organized into nine categories: Ableism, Fatphobia, Religious Intolerance, LGBTQ+ Phobia, Misogyny, Xenophobia, Racism, Hetarism, and Political Prejudice[2].

3.3 Data Selection

We applied an inclusion filter based on the keywords present in the PrejudicePT-br dictionary. This process selected messages containing at least one of the prejudiced words cataloged in PrejudicePT-br. It is important to note that the presence of a word from the PrejudicePT-br dictionary in a message does not necessarily imply that the message conveys prejudice, as the meaning of prejudiced words depends on context. After applying the inclusion filter, the WhatsApp dataset was reduced from 813,106 unique messages to 37,531, while the Telegram dataset was reduced from 767,847 unique messages to 58,787. Even after this reduction, the number of messages remains too large for manual annotation.

3.4 Data Anonymization

To ensure user privacy, personal data such as names and phone numbers were anonymized. Additionally, we applied a hash function to generate a unique and anonymous identifier for each user based on their phone number. Furthermore, a hash function was also used to create a unique and anonymous identifier for each group, derived from its name. Since these groups are publicly accessible, our approach does not violate WhatsApp's privacy policy[3] nor the General Data Protection Law (LGPD).

[1] https://faroldigital.info/

Table 2: Basic Statistics of the PrejudiceWhatsApp.Br and PrejudiceTelegram.Br Datasets.

Metric	PrejudiceWhatsApp.Br	PrejudiceTelegram.Br
Number of Unique Messages	3.000	3.000
Mean and Standard Deviation of the Number of Tokens	34,72 $\pm$ 30,76	50, 81 $\pm$ 40,03
Minimum Number of Tokens	1	1
Maximum Number of Tokens	186	183

3.5 Data Labeling

Data labeling is a complex challenge, as it requires determining whether a given text is prejudiced or not, meaning whether it is related to any form of prejudice. Next, we describe the manual labeling process applied to the textual content of the messages obtained after the keyword filtering step. It is important to emphasize that the labeling process was entirely manual to ensure that the textual corpus is of high quality. Three annotators conducted the labeling process, and disagreements were resolved through a collective review to ensure consistency and reliability.

Since the number of messages obtained after the data filtering step remains significantly large (37,531 for WhatsApp and 58,787 for Telegram), and to ensure balanced datasets, we applied the following strategy: We constructed two datasets, named PrejudiceWhatsApp.Br and PrejudiceTelegram.Br, each containing 3,000 messages that circulated on WhatsApp and Telegram, respectively. Each dataset consists of 1,500 unique messages labeled as prejudiced and 1,500 unique messages labeled as non-prejudiced [4].

To obtain these datasets, we follow the approach described next. We randomly selected messages from the filtered dataset and manually labeled each one. This process was repeated until we obtained 1,500 messages labeled as prejudiced (label 1) and 1,500 messages labeled as non-prejudiced (label 0) for each platform.

In Table 2, we present basic statistical metrics computed for the PrejudiceWhatsApp.Br and PrejudiceTelegram.Br datasets, including the number of unique messages, the minimum and maximum number of tokens, as well as the mean and standard deviation of the token count. From Table 2, it can be observed that the characteristics of the messages composing the PrejudiceWhatsApp.Br and PrejudiceTelegram.Br datasets do not appear to change substantially, suggesting similar text structures across both platforms.

[2]The PrejudicePT-br dictionary is available at https://github.com/jmmfilho/PrejudicePT-br

[3]https://www.whatsapp.com/legal/privacy-policy

[4]https://github.com/jmmfilho/PrejudicePT-br

4 EXPERIMENTAL EVALUATION

4.1 Baseline Evaluation

To provide a baseline for the prejudice detection problem in Portuguese text messages from Telegram and WhatsApp, a series of experiments was conducted using the PrejudiceWhatsApp.Br and PrejudiceTelegram.Br datasets.

4.1.1 Features and Classification Algorithms

As previously mentioned, two distinct feature extraction methods were evaluated: BoW and TF-IDF. Pretrained embedding vectors were not used due to the high occurrence of misspelled words, emoticons, and neologisms in the corpus. In this context, the BoW and TF-IDF methods were chosen due to their simplicity, speed, and widespread use in text classification tasks.

Before applying the BoW and TF-IDF methods, the text was converted to lowercase. It is important to note that emojis are highly prevalent in the dataset and play a significant role in the language used in instant messaging applications. For this reason, they were retained in the preprocessing step. However, since emoji combinations can generate different types of tokens, a whitespace separation strategy was applied, ensuring that each emoji is treated as an individual token. Additionally, URL normalization was performed, where only the domain name was preserved. Due to the lexical diversity of the corpus, the resulting feature vectors are sparse and exhibit high dimensionality.

Three different tokenization strategies were evaluated: unigrams, bigrams, and trigrams. While this approach results in high-dimensional vectors, it is expected to reveal distinct patterns in the messages, as bigrams and trigrams can capture more context-related information. Additionally, to assess the impact of preprocessing techniques, two approaches were considered: i) no preprocessing, and ii) using stop-words removal and lemmatization. These techniques aim to reduce noise, enabling a more precise representation of the relevant features present in the messages.

Thus, 12 different execution scenarios were created by combining two feature extraction methods (BoW and TF-IDF), three tokenization strategies (unigrams, bigrams, and trigrams), and two preprocessing approaches (with and without preprocessing). For each of these scenarios, we evaluated nine classical classification algorithms (Pranckevičius and Marcinkevičius, 2017), covering different categories: linear models (Logistic Regression - LR), generative models (Bernoulli Naive Bayes - BNB and Multinomial Naive Bayes - MNB), instance-based learning (K-Nearest Neighbors - KNN), support vector machines (Linear Support Vector Machine - LSVM and Stochastic Gradient Descent - SGD), ensemble learning methods (Random Forest - RF and Gradient Boosting - GB), and neural networks (Multilayer Perceptron - MLP). As a result, for each developed dataset, a total of 108 experiments was conducted. The classification algorithms were implemented using the scikit-learn Python library (Pedregosa et al., 2011). In this study, due to the infrastructure available, we chose to evaluate only classical machine learning algorithms, leaving the investigation of popular language models (e.g., BERTimbau, GPT-4, and DeepSeek V3) for future work.

For the MLP method, a batch size of 64 was used, along with an early stopping strategy. In this approach, 10% of the training data was reserved for validation, and training was halted if the validation performance did not improve by at least 0.001 for five consecutive epochs. All other hyperparameters were kept at their default values for all classification algorithms. Although a systematic hyperparameter optimization was not performed, the diversity of tested approaches provides valuable insights into which learning strategies may be more suitable for the investigated problem, thereby establishing a baseline. All data and code used in the experiments are available in our online repository[5].

4.1.2 Performance Metrics

As previously mentioned, the investigated problem is a binary classification task, where prejudice represents the positive class (which is also our class of interest), and non-prejudice represents the negative class. To evaluate the performance of each model, the following metrics were used: False Positive Rate (FPR), Precision (PRE), Recall (REC), and F1-score (F1). Since we applied k-fold cross-validation ($k = 5$), we will report the mean and standard deviation of each metric across all conducted experiments.

After performing cross-validation, we selected the best classifier and the most effective features. Next, we retrained the model using a randomly selected training set, which corresponds to 80% of the total available data. Subsequently, we evaluated the model's performance using the remaining 20% of the data, which was initially set aside to form the test set.

4.2 Using the PrejudicePT-br Dictionary

In this series of experiments, we aimed to enhance predictive models by combining the BoW and TF-IDF feature extraction methods with the categories present in the PrejudicePT-br dictionary. For each message, we computed the number of words belonging to each of the nine categories in PrejudicePT-br. As a result, nine additional features were incorporated into the experiments. Finally, we evaluated the strategy of using the words from the PrejudicePT-br dictionary as features, by calculating how many times each word appears in a given message. This approach resulted in a feature set of 842 attributes.

5 RESULTS

In this section, we present and discuss the results obtained for both datasets, PrejudiceWhatsApp.Br and PrejudiceTelegram.Br, in both the baseline evaluation and the experiments incorporating the PrejudicePT-br dictionary.

5.1 Baseline Evaluation Results

Each of the tables presented below contains information on six different scenarios. For each scenario, we highlight the values of the performance metrics (AUC Score, Precision, Recall, and F1-score) obtained by each of the nine evaluated classifiers, along with the number of generated features and the training time required for the classification model. More specifically, the six evaluated scenarios were as follows: (a) Unigrams only, without preprocessing, (b) Unigrams only, with stopword removal and lemmatization, (c) Unigrams and bigrams, without preprocessing, (d) Unigrams and bigrams, with stopword removal and lemmatization, (e) Unigrams, bigrams, and trigrams, without preprocessing, and (f) Unigrams, bigrams, and trigrams, with stopword removal and lemmatization.

[5]https://github.com/jmmfilho/PrejudicePT-br

Table 3: Baseline Results on PrejudiceTelegram.Br Using the BoW Method.

(a) BoW-1. Features: 14.491, Time: 313.0s.

Method	Auc Score	Precision	Recall	F1-score
LR	0.90	0.82±0.02	0.81±0.01	0.820±0.01
BNB	0.85	0.64±0.02	0.90±0.02	0.752±0.01
MNB	0.87	0.73±0.01	0.89±0.02	0.808±0.01
LSVM	0.89	0.82±0.02	0.81±0.01	0.820±0.01
KNN	0.70	0.59±0.02	0.80±0.01	0.686±0.01
SGD	0.89	0.83±0.04	0.80±0.05	0.812±0.01
RF	0.90	0.83±0.02	0.79±0.02	0.813±0.01
GB	0.93	0.92±0.01	0.80±0.01	**0.860±0.00**
MLP	0.87	0.80±0.02	0.79±0.03	0.796±0.02

(b) BoW-1 W/Pre. Features: 15.218, Time: 141.3s.

Method	Auc Score	Precision	Recall	F1-score
LR	0.90	0.82±0.02	0.81±0.01	0.818±0.01
BNB	0.85	0.63±0.01	0.92±0.01	0.755±0.01
MNB	0.86	0.73±0.02	0.89±0.01	0.808±0.01
LSVM	0.88	0.81±0.03	0.79±0.01	0.801±0.02
KNN	0.71	0.64±0.01	0.71±0.03	0.680±0.02
SGD	0.87	0.81±0.02	0.77±0.02	0.794±0.01
RF	0.90	0.84±0.01	0.78±0.01	0.812±0.00
GB	0.91	0.89±0.03	0.78±0.03	**0.834±0.01**
MLP	0.87	0.79±0.03	0.81±0.01	0.805±0.01

(c) BoW-1,2. Features: 102.297, Time: 872.5s.

Method	Auc Score	Precision	Recall	F1-score
LR	0.90	0.82±0.02	0.81±0.01	0.822±0.01
BNB	0.85	0.59±0.01	0.96±0.00	0.735±0.01
MNB	0.86	0.72±0.02	0.89±0.02	0.801±0.01
LSVM	0.90	0.83±0.01	0.81±0.01	0.825±0.00
KNN	0.60	0.50±0.00	0.98±0.00	0.670±0.00
SGD	0.88	0.81±0.02	0.79±0.03	0.804±0.02
RF	0.87	0.79±0.03	0.78±0.01	0.789±0.01
GB	0.93	0.92±0.01	0.80±0.01	**0.859±0.00**
MLP	0.88	0.78±0.02	0.81±0.02	0.801±0.02

(d) BoW-1,2 W/Pre. features: 70.176, Time: 1647.4s.

Method	Auc Score	Precision	Recall	F1-score
LR	0.89	0.80±0.02	0.81±0.01	0.813±0.01
BNB	0.85	0.58±0.00	0.97±0.01	0.727±0.00
MNB	0.84	0.70±0.01	0.91±0.01	0.793±0.01
LSVM	0.89	0.81±0.02	0.80±0.01	0.806±0.01
KNN	0.59	0.51±0.01	0.94±0.03	0.668±0.00
SGD	0.87	0.79±0.01	0.77±0.01	0.784±0.01
RF	0.87	0.78±0.04	0.78±0.02	0.784±0.02
GB	0.91	0.91±0.01	0.75±0.01	**0.826±0.00**
MLP	0.87	0.79±0.02	0.80±0.01	0.798±0.01

(e) BoW-1,2,3. Features: 222.563, Time: 2134.5s.

Method	Auc Score	Precision	Recall	F1-score
LR	0.89	0.81±0.02	0.80±0.02	0.809±0.01
BNB	0.83	0.56±0.00	0.98±0.00	0.715± 0.00
MNB	0.83	0.70±0.01	0.91±0.01	0.794±0.01
LSVM	0.89	0.82±0.01	0.79±0.00	0.812±0.01
KNN	0.56	0.50±0.00	0.99±0.00	0.669±0.00
SGD	0.87	0.80±0.02	0.79±0.04	0.795±0.02
RF	0.86	0.75±0.02	0.80±0.02	0.780±0.02
GB	0.93	0.92±0.01	0.79±0.01	**0.857±0.00**
MLP	0.86	0.76±0.01	0.82±0.04	0.793±0.02

(f) BoW-1,2,3 W/Pre. Features: 164.396, Time: 1458.8s.

Method	Auc Score	Precision	Recall	F1-score
LR	0.89	0.81±0.02	0.80±0.02	0.806±0.01
BNB	0.84	0.54±0.00	0.98±0.01	0.706± 0.00
MNB	0.83	0.68±0.00	0.93±0.01	0.787±0.00
LSVM	0.89	0.82±0.01	0.79±0.02	0.808±0.00
KNN	0.59	0.51±0.01	0.95±0.00	0.669±0.00
SGD	0.87	0.80±0.02	0.79±0.02	0.798±0.01
RF	0.86	0.77±0.03	0.79±0.03	0.784±0.02
GB	0.91	0.90±0.03	0.77±0.01	**0.830±0.00**
MLP	0.86	0.75±0.02	0.84±0.03	0.798±0.01

5.1.1 Results Obtained for Telegram

The results obtained from the experiments conducted with the PrejudiceTelegram.Br dataset are summarized in Tables 3 and 4. Table 3 presents the results for the BoW (Bag of Words) feature extraction method. Table 4 illustrates the results for the TF-IDF feature extraction method. From Tables 3 and 4, we observe that the Gradient Boosting (GB) and Support Vector Machine (SVM) classifiers generally achieved the best performance, considering the F1-score. On the other hand, Bernoulli Naive Bayes (BNB) and K-Nearest Neighbors (KNN) had the lowest average performance. The remaining classifiers performed consistently well across all scenarios. Additionally, we note that the BoW and TF-IDF strategies yielded similar results, indicating that both methods are viable for prejudiced language detection in the Telegram dataset.

5.1.2 Results Obtained for WhatsApp

The results obtained from the experiments conducted with the PrejudiceWhatsApp.Br dataset are summarized in Tables 5 and 6. Table 5 presents the results for the BoW feature extraction method. Table 6 illustrates the results for the TF-IDF feature extraction method. From Tables 5 and 6, we observe that the Logistic Regression (LR), Support Vector Machine (SVM) and Gradient Boosting (GB) classifiers generally achieved the best performance in terms of F1-score. On the other hand, the K-Nearest Neighbors (KNN) classifier consistently demonstrated the worst performance across all tested configurations. This suggests that KNN struggled with the classification task in this context. The remaining classifiers performed consistently well across all scenarios, further reinforcing the effectiveness of different machine learning methods for prejudiced language detection on WhatsApp.

Table 4: Baseline Results on PrejudiceTelegram.Br Using the TF-IDF Method.

(a) TF-IDF-1. Features: 16.371, Time: 238.3s.

Method	Auc Score	Precision	Recall	F1-score
LR	0.85	0.78±0.02	0.73±0.02	0.763±0.02
BNB	0.85	0.64±0.02	0.90±0.02	0.752±0.01
MNB	0.85	0.75±0.01	0.79±0.02	0.773±0.01
LSVM	0.89	0.81±0.01	0.79±0.02	0.807±0.01
KNN	0.78	0.74±0.01	0.69±0.07	0.716±0.00
SGD	0.89	0.80±0.02	0.79±0.06	0.795±0.02
RF	0.88	0.81±0.03	0.77±0.02	0.796±0.01
GB	0.93	0.91±0.01	0.79±0.00	**0.853±0.00**
MLP	0.86	0.78±0.02	0.77±0.02	0.776±0.02

(b) TF-IDF-1 W/Pre. Features: 15.218, Time: 164.1s.

Method	Auc Score	Precision	Recall	F1-score
LR	0.85	0.78±0.01	0.74±0.01	0.762±0.00
BNB	0.85	0.63±0.01	0.92±0.01	0.755±0.01
MNB	0.85	0.75±0.01	0.80±0.02	0.773±0.01
LSVM	0.89	0.81±0.02	0.79±0.02	0.808±0.02
KNN	0.78	0.73±0.01	0.69±0.02	0.713±0.00
SGD	0.88	0.81±0.03	0.78±0.03	0.796±0.02
RF	0.89	0.82±0.01	0.77±0.01	0.800±0.01
GB	0.91	0.88±0.02	0.76±0.02	**0.820±0.01**
MLP	0.86	0.77±0.01	0.78±0.03	0.779±0.01

(c) TF-IDF-1,2. Features: 102.297, Time: 954.8s.

Method	Auc Score	Precision	Recall	F1-score
LR	0.80	0.76±0.02	0.67±0.02	0.713±0.02
BNB	0.85	0.59±0.01	0.96±0.00	0.735±0.01
MNB	0.83	0.76±0.01	0.74±0.04	0.753±0.02
LSVM	0.87	0.80±0.02	0.76±0.01	0.783±0.01
KNN	0.79	0.74±0.01	0.68±0.01	0.713±0.00
SGD	0.88	0.78±0.00	0.79±0.03	0.793±0.01
RF	0.87	0.78±0.01	0.78±0.03	0.783±0.02
GB	0.93	0.91±0.00	0.79±0.01	**0.855±0.00**
MLP	0.84	0.76±0.03	0.73±0.06	0.746±0.03

(d) TF-IDF-1,2 W/Pre. Features: 84.810, Time: 951.3s.

Method	Auc Score	Precision	Recall	F1-score
LR	0.80	0.76±0.02	0.66±0.02	0.711±0.01
BNB	0.85	0.58±0.00	0.97±0.01	0.729±0.00
MNB	0.83	0.74±0.01	0.75±0.02	0.750±0.01
LSVM	0.87	0.80±0.02	0.76±0.01	0.787±0.00
KNN	0.78	0.72±0.01	0.69±0.03	0.708±0.01
SGD	0.88	0.78±0.02	0.81±0.03	0.799±0.01
RF	0.87	0.79±0.01	0.78±0.01	0.790±0.01
GB	0.91	0.89±0.02	0.76±0.01	**0.820±0.01**
MLP	0.85	0.76±0.01	0.77±0.02	0.768±0.01

(e) TF-IDF-1,2,3. Features: 222.563, Time: 2231.7s.

Method	Auc Score	Precision	Recall	F1-score
LR	0.78	0.76±0.03	0.62±0.02	0.687±0.02
BNB	0.83	0.56±0.00	0.98±0.00	0.715±0.00
MNB	0.81	0.75±0.01	0.71±0.04	0.734±0.02
LSVM	0.85	0.78±0.02	0.73±0.01	0.755±0.01
KNN	0.79	0.73±0.01	0.67±0.02	0.703±0.01
SGD	0.86	0.80±0.03	0.73±0.05	0.761±0.02
RF	0.85	0.75±0.01	0.79±0.04	0.776±0.01
GB	0.93	0.91±0.01	0.79±0.01	**0.851±0.00**
MLP	0.83	0.74±0.02	0.74±0.04	0.745±0.02

(f) TF-IDF-1,2,3 W/Pre. features: 164.396, Time: 2193.2s.

Method	Auc Score	Precision	Recall	F1-score
LR	0.78	0.75±0.02	0.62±0.02	0.683±0.02
BNB	0.84	0.54±0.00	0.98±0.00	0.706±0.00
MNB	0.81	0.73±0.00	0.73±0.03	0.735±0.01
LSVM	0.85	0.79±0.01	0.73±0.02	0.764±0.01
KNN	0.78	0.72±0.01	0.68±0.02	0.705±0.01
SGD	0.85	0.79±0.02	0.72±0.06	0.755±0.02
RF	0.86	0.76±0.03	0.79±0.01	0.777±0.02
GB	0.91	0.88±0.02	0.76±0.02	**0.818±0.01**
MLP	0.83	0.75±0.01	0.77±0.02	0.760±0.01

5.2 Results from Using the PrejudicePT-br Dictionary

Initially, we evaluated the strategy of combining the nine categories from the PrejudicePT-br dictionary with the BoW and TF-IDF feature extraction methods, while also incorporating unigrams, bigrams, and trigrams, along with text preprocessing. The results obtained using this strategy are presented in Table 7.

Notably, the best baseline result for PrejudiceTelegram.Br, with an F1-score of 0.860, was obtained using Gradient Boosting (GB) with unigrams and no preprocessing. However, the best result obtained using the dictionary-based approach (incorporating the categories from PrejudicePT-br) for PrejudiceTelegram.Br reached an F1-score of 0.857. This was achieved using GB, with unigrams, bigrams, and trigrams, along without preprocessing, and the BoW method.

It is worth noting that the best baseline result for PrejudiceWhatsApp.Br, with an F1-score of 0.868, was obtained using Gradient Boosting (GB) with unigrams and bigrams, without text preprocessing, and using the TF-IDF method. In contrast, the best result using the dictionary-based approach (incorporating the categories from PrejudicePT-br) for PrejudiceWhatsApp.Br achieved an F1-score of 0.866. This result was obtained using GB with unigrams, bigrams, and trigrams, without preprocessing, and the TF-IDF method.

Next, we conducted experiments to evaluate the use of the nine categories from the PrejudicePT-br dictionary as standalone features, without combining them with the BoW or TF-IDF feature extraction methods. The results of these experiments are shown in Table 8. Note that, the best result for PrejudiceTelegram.Br achieved an F1-score of 0.640, using only 9 features and a training time of 5.8s. This per-

Table 5: Baseline Results on PrejudiceWhatsApp.Br Using the BoW Method.

(a) BoW-1. Features: 10.924, Time: 319.2s.

Method	Auc Score	Precision	Recall	F1-score
LR	0.93	0.88±0.00	0.84±0.01	**0.863 ± 0.00**
BNB	0.90	0.70±0.01	0.92±0.01	0.801±0.01
MNB	0.91	0.77±0.00	0.90±0.01	0.834±0.01
LSVM	0.92	0.87±0.01	0.83±0.00	0.855±0.00
KNN	0.76	0.81±0.05	0.44±0.07	0.566±0.05
SGD	0.92	0.85±0.01	0.83±0.02	0.844±0.00
RF	0.92	0.87±0.02	0.81±0.01	0.836±0.00
GB	0.94	0.94±0.01	0.79±0.02	0.862±0.01
MLP	0.90	0.84±0.03	0.83±0.02	0.846±0.00

(b) BoW-1 W/Pre. features: 9.820, Time: 308.3s.

Method	Auc Score	Precision	Recall	F1-score
LR	0.92	0.86±0.01	0.82±0.01	**0.841 ± 0.01**
BNB	0.90	0.69±0.01	0.92±0.01	0.793±0.00
MNB	0.91	0.77±0.01	0.89±0.01	0.830±0.00
LSVM	0.91	0.85±0.00	0.82±0.01	0.837±0.00
KNN	0.77	0.81±0.06	0.52±0.08	0.629±0.04
SGD	0.90	0.86±0.01	0.80±0.02	0.835±0.01
RF	0.92	0.86±0.02	0.79±0.02	0.831±0.00
GB	0.93	0.94±0.01	0.76±0.02	0.840±0.02
MLP	0.91	0.84±0.02	0.82±0.01	0.837±0.01

(c) BoW-1,2. Features: 60.344, Time: 1504.7s.

Method	Auc Score	Precision	Recall	F1-score
LR	0.92	0.86±0.01	0.84±0.00	0.850±0.00
BNB	0.89	0.65±0.01	0.93±0.01	0.773±0.01
MNB	0.89	0.76±0.01	0.88±0.01	0.823±0.01
LSVM	0.92	0.87±0.01	0.83±0.01	0.853±0.01
KNN	0.69	0.85±0.11	0.28±0.2	0.372±0.02
SGD	0.91	0.85±0.01	0.82±0.01	0.835± 0.01
RF	0.91	0.82±0.03	0.83±0.01	0.829±0.01
GB	0.94	0.94±0.01	0.78±0.02	**0.860±0.01**
MLP	0.90	0.84±0.01	0.83±0.01	0.838±0.01

(d) BoW-1,2 W/Pre. Features: 48.576, Time: 1421.5s.

Method	Auc Score	Precision	Recall	F1-score
LR	0.92	0.85±0.01	0.82±0.01	0.837±0.00
BNB	0.89	0.63±0.00	0.94±0.01	0.761±0.00
MNB	0.89	0.76±0.01	0.89±0.01	0.822±0.01
LSVM	0.92	0.87±0.00	0.81±0.00	0.841±0.00
KNN	0.67	0.75±0.07	0.38±0.17	0.484±0.11
SGD	0.91	0.86±0.02	0.80±0.01	0.834± 0.00
RF	0.90	0.83±0.03	0.81±0.01	0.827±0.01
GB	0.93	0.94±0.01	0.75±0.02	**0.842±0.01**
MLP	0.90	0.83±0.03	0.82±0.01	0.828±0.01

(e) BoW-1,2,3. Features: 125.729, Time: 3716.1s.

Method	Auc Score	Precision	Recall	F1-score
LR	0.91	0.84±0.02	0.83±0.00	0.844±0.01
BNB	0.88	0.61±0.00	0.94±0.00	0.748± 0.00
MNB	0.88	0.75±0.01	0.90±0.01	0.820±0.01
LSVM	0.92	0.86±0.02	0.83±0.00	0.847 ±0.01
KNN	0.64	0.84±0.15	0.30±0.34	0.304±0.30
SGD	0.91	0.84±0.02	0.84±0.01	0.836±0.00
RF	0.89	0.78±0.01	0.84±0.01	0.810±0.01
GB	0.94	0.94±0.01	0.78±0.02	**0.858±0.02**
MLP	0.90	0.83±0.01	0.84±0.01	0.838±0.00

(f) BoW-1,2,3 W/Pre. Features: 113.634, Time: 1150.4s.

Method	Auc Score	Precision	Recall	F1-score
LR	0.92	0.87±0.01	0.82±0.00	0.848±0.01
BNB	0.88	0.60±0.00	0.96±0.01	0.743± 0.00
MNB	0.89	0.75±0.01	0.90±0.01	0.822±0.01
LSVM	0.92	0.87±0.01	0.83±0.00	**0.853 ±0.00**
KNN	0.70	0.87±0.10	0.28±0.22	0.374±0.17
SGD	0.91	0.86±0.02	0.82±0.01	0.842±0.00
RF	0.91	0.83±0.03	0.80±0.02	0.822±0.01
GB	0.93	0.92±0.00	0.78±0.02	0.846±0.01
MLP	0.91	0.85±0.02	0.83±0.02	0.842±0.01

formance was lower than the one obtained with the combined approach (PrejudicePT-br + BoW), which reached an F1-score of 0.857. However, the combined approach used 222,563 features and required 2,134.6s for training. For the PrejudiceWhatsApp.Br dataset, using only the PrejudicePT-br dictionary as features resulted in an F1-score of 0.733, with just 9 features and a training time of 5.7s. This performance was also lower than that of the combined approach (PrejudicePT-br + TF-IDF), which obtained an F1-score of 0.866. However, the combined approach used 153,813 features and required 1,603.3s for training. These results suggest that for specific scenarios requiring frequent model retraining, a viable alternative could be to use only the nine categories from the PrejudicePT-br dictionary as features, significantly reducing computational cost while maintaining reasonable classification performance.

Finally, we conducted experiments to evaluate the use of the 842 words from the PrejudicePT-br dictionary as standalone features, without combining them with the BoW or TF-IDF feature extraction methods. The results of these experiments are shown in Table 9.

In this final experiment, the best result for PrejudiceTelegram.Br achieved an F1-score of 0.713, using MNB and only 842 features, with a training time of 24.8s. This performance was lower than the one obtained with the combined approach (PrejudicePT-br + BoW + GB), which reached an F1-score of 0.857. However, the combined approach used 222.563 features and required 2,134.6s for training. For the PrejudiceWhatsApp.Br dataset, using only the words of PrejudicePT-br dictionary as features resulted in an F1-score of 0.812, with just 842 features and a training time of 34.71s. This performance was also lower than that of the combined approach (PrejudicePT-br + TF-IDF + GB), which obtained an F1-score of 0.866. However, the combined approach used 153,813 features and required 1,603.3s for training. Besides, the best result for PrejudiceWhatsApp.Br achieved an F1-score of 0.812, a lower value compared with the F1-

Table 6: Baseline Results on PrejudiceWhatsApp.Br Using the TF-IDF Method.

(a) TF-IDF-1. Features: 12.549, Time: 130.0s.

Method	Auc Score	Precision	Recall	F1-score
LR	0.89	0.83±0.02	0.78±0.01	0.810±0.01
BNB	0.90	0.71±0.01	0.91±0.01	0.804±0.01
MNB	0.88	0.79±0.01	0.81±0.02	0.803±0.01
LSVM	0.92	0.88±0.01	0.83±0.01	0.855±0.01
KNN	0.87	0.81±0.03	0.78±0.01	0.800±0.01
SGD	0.92	0.86±0.02	0.83±0.01	0.848±0.01
RF	0.92	0.86±0.02	0.81±0.01	0.837±0.00
GB	0.94	0.92±0.00	0.80±0.02	**0.861±0.01**
MLP	0.91	0.85±0.01	0.82±0.01	0.841±0.00

(b) TF-IDF-1 W/Pre. Features: 11.344, Time: 146.1s.

Method	Auc Score	Precision	Recall	F1-score
LR	0.89	0.84±0.02	0.78±0.01	0.809±0.01
BNB	0.90	0.70±0.01	0.91±0.01	0.800±0.00
MNB	0.88	0.80±0.01	0.81±0.02	0.806±0.01
LSVM	0.92	0.87±0.02	0.82±0.01	0.850±0.01
KNN	0.86	0.81±0.02	0.78±0.02	0.797±0.02
SGD	0.91	0.86±0.01	0.82±0.02	0.844±0.01
RF	0.92	0.86±0.01	0.79±0.01	0.831±0.00
GB	0.93	0.92±0.01	0.78±0.02	**0.850±0.01**
MLP	0.90	0.84±0.02	0.82±0.01	0.831±0.01

(c) TF-IDF-1,2. Features: 72.560, Time: 944.0s.

Method	Auc Score	Precision	Recall	F1-score
LR	0.85	0.82±0.02	0.74±0.02	0.779±0.02
BNB	0.89	0.66±0.01	0.94±0.01	0.781±0.01
MNB	0.86	0.79±0.02	0.77±0.02	0.786±0.01
LSVM	0.90	0.85±0.02	0.79±0.01	0.826±0.01
KNN	0.86	0.80±0.02	0.77±0.01	0.791±0.01
SGD	0.90	0.85±0.02	0.80±0.01	0.831±0.00
RF	0.91	0.84±0.01	0.82±0.01	0.833±0.01
GB	0.94	0.92±0.01	0.81±0.02	**0.868±0.01**
MLP	0.89	0.83±0.03	0.81±0.01	0.825±0.01

(d) TF-IDF-1,2 W/Pre. Features: 59.134, Time: 1012.8s.

Method	Auc Score	Precision	Recall	F1-score
LR	0.85	0.82±0.02	0.74±0.02	0.782±0.01
BNB	0.89	0.64±0.00	0.95±0.01	0.770±0.00
MNB	0.85	0.79±0.01	0.77±0.03	0.784±0.02
LSVM	0.90	0.86±0.02	0.79±0.01	0.828±0.01
KNN	0.86	0.81±0.02	0.78±0.02	0.798±0.02
SGD	0.90	0.86±0.02	0.79±0.02	0.828±0.01
RF	0.90	0.83±0.03	0.81±0.02	0.827±0.01
GB	0.93	0.92±0.00	0.79±0.01	**0.853±0.01**
MLP	0.90	0.83±0.02	0.80±0.01	0.821±0.00

(e) TF-IDF-1,2,3. Features: 153.813, Time: 1603.3s.

Method	Auc Score	Precision	Recall	F1-score
LR	0.84	0.81±0.03	0.71±0.02	0.761±0.02
BNB	0.89	0.62±0.00	0.94±0.11	0.751±0.00
MNB	0.84	0.79±0.02	0.76±0.01	0.780±0.01
LSVM	0.89	0.84±0.02	0.77±0.02	0.810±0.01
KNN	0.86	0.80±0.01	0.77±0.01	0.789±0.01
SGD	0.89	0.84±0.02	0.80±0.01	0.819±0.01
RF	0.90	0.81±0.02	0.82±0.02	0.818±0.01
GB	0.94	0.92±0.01	0.81±0.01	**0.866±0.00**
MLP	0.88	0.81±0.04	0.80±0.00	0.806±0.02

(f) TF-IDF-1,2,3 W/Pre. Features: 113.634, Time: 1451.7s.

Method	Auc Score	Precision	Recall	F1-score
LR	0.84	0.81±0.03	0.72±0.02	0.765±0.01
BNB	0.88	0.60±0.00	0.96±0.01	0.743±0.00
MNB	0.84	0.78±0.01	0.77±0.03	0.780±0.01
LSVM	0.89	0.85±0.02	0.77±0.01	0.812±0.01
KNN	0.86	0.81±0.02	0.78±0.02	0.797±0.02
SGD	0.89	0.84±0.02	0.78±0.01	0.815±0.01
RF	0.90	0.81±0.03	0.80±0.03	0.811±0.02
GB	0.93	0.91±0.00	0.79±0.02	**0.851±0.01**
MLP	0.88	0.82±0.03	0.78±0.02	0.808±0.02

score of 0.868 obtained by the combined approach.

These results suggest that for specific scenarios requiring frequent model retraining, a viable alternative could be to use only the 842 words from the PrejudicePT-br dictionary as features, significantly reducing computational cost while maintaining reasonable classification performance.

5.3 Threats to Validity

During the execution of the experiments, several threats to validity were identified (Wohlin et al., 2012). Next, we discuss some of them. The message collection occurred during a period of intense political debate, which may have increased the number of prejudiced messages, potentially affecting the distribution of the dataset. The messages were collected from 179 public WhatsApp groups and 150 public Telegram groups/channels, primarily focused on political debates. This sample may not fully capture the general behavior of groups in Brazil.

6 CONCLUSIONS AND FUTURE WORK

In this study, we introduced two datasets, named PrejudiceWhatsApp.Br and PrejudiceTelegram.Br, containing prejudiced messages in Brazilian Portuguese (PT-BR) that circulated in public WhatsApp and Telegram groups, respectively. Additionally, we developed a dictionary of prejudiced words for Brazilian Portuguese, named PrejudicePT-br, consisting of 842 words organized into nine categories. Based on the PrejudicePT-br dictionary, we proposed an approach for the automatic detection of prejudiced messages. Finally, a series of experiments was conducted to evaluate the proposed approach, which achieved a best F1-score of 0.868. This result demonstrates the feasibility of the proposed method. As future

Table 7: Results from Using the 9 Categories of the PrejudicePT-br Dictionary with BoW and TF-IDF.

(a) PrejudicePT-br + BoW-1,2,3. Features: 153.813, Time: 1446.0s - PrejudiceWhatsApp.Br.

Method	Auc Score	Precision	Recall	F1-score
LR	0.91	0.84±0.02	0.83±0.00	0.851±0.01
BNB	0.88	0.61±0.00	0.94±0.00	0.751±0.00
MNB	0.88	0.75±0.01	0.90±0.01	0.823±0.01
LSVM	0.92	0.86±0.02	0.83±0.00	**0.860±0.01**
KNN	0.64	0.84±0.15	0.30±0.34	0.372±0.30
SGD	0.91	0.84±0.02	0.84±0.01	0.842±0.01
RF	0.89	0.78±0.01	0.84±0.01	0.822±0.00
GB	0.94	0.94±0.01	0.78±0.02	0.858±0.01
MLP	0.90	0.83±0.01	0.84±0.01	0.837±0.00

(b) PrejudicePT-br + BoW-1,2,3. Features: 222.563, Time: 2134.6s - PrejudiceTelegram.Br.

Method	Auc Score	Precision	Recall	F1-score
LR	0.89	0.80±0.02	0.81±0.02	0.809±0.02
BNB	0.83	0.56±0.00	0.98±0.00	0.715±0.00
MNB	0.83	0.68±0.01	0.93±0.01	0.794±0.00
LSVM	0.89	0.81±0.03	0.81±0.01	0.812±0.01
KNN	0.54	0.50±0.00	0.99±0.00	0.669±0.00
SGD	0.87	0.80±0.01	0.78±0.04	0.795±0.02
RF	0.85	0.74±0.03	0.81±0.02	0.780±0.03
GB	0.93	0.93±0.01	0.78±0.01	**0.857±0.00**
MLP	0.86	0.79±0.03	0.82±0.02	0.793± 0.01

(c) PrejudicePT-br + TF-IDF-1,2,3. Features: 153.813, Time: 1603.3s - PrejudiceWhatsApp.Br.

Method	Auc Score	Precision	Recall	F1-score
LR	0.87	0.77±0.02	0.85±0.01	0.761±0.02
BNB	0.88	0.61±0.00	0.94±0.00	0.751±0.00
MNB	0.86	0.80±0.02	0.83±0.01	0.780±0.01
LSVM	0.90	0.81±0.03	0.83±0.00	0.810±0.01
KNN	0.79	0.74±0.02	0.71±0.02	0.789±0.01
SGD	0.90	0.83±0.03	0.82±0.01	0.819±0.01
RF	0.88	0.77±0.03	0.83±0.01	0.818±0.01
GB	0.94	0.94±0.01	0.79±0.02	**0.866±0.00**
MLP	0.87	0.78±0.03	0.82±0.01	0.806±0.02

(d) PrejudicePT-br + TF-IDF-1,2,3. Features: 222.563, Time: 2231.8s - PrejudiceTelegram.Br.

Method	Auc Score	Precision	Recall	F1-score
LR	0.85	0.76±0.02	0.81±0.03	0.687±0.02
BNB	0.83	0.56±0.00	0.98±0.00	0.715±0.00
MNB	0.85	0.77±0.02	0.79±0.04	0.734±0.02
LSVM	0.87	0.78±0.02	0.81±0.02	0.755±0.01
KNN	0.72	0.70±0.02	0.54±0.05	0.703±0.01
SGD	0.88	0.79±0.02	0.80±0.03	0.761±0.02
RF	0.85	0.74±0.02	0.79±0.03	0.776±0.01
GB	0.93	0.92±0.01	0.79±0.02	**0.851±0.00**
MLP	0.84	0.76±0.03	0.76±0.08	0.745±0.02

Table 8: Results from Using the 9 Categories of the PrejudicePT-br Dictionary.

(a) PrejudicePT-br. Features: 9, Time: 5.7s - PrejudiceWhatsApp.Br

Method	Auc Score	Precision	Recall	F1-score
LR	0.77	0.79±0.03	0.67±0.03	0.730±0.03
BNB	0.73	0.63±0.01	0.82±0.01	0.719±0.00
MNB	0.72	0.75±0.03	0.49±0.01	0.594±0.07
LSVM	0.77	0.79±0.03	0.66±0.03	0.726±0.03
KNN	0.74	0.72±0.01	0.65±0.15	0.664±0.08
SGD	0.75	0.78±0.03	0.65±0.04	0.714±0.03
RF	0.79	0.80±0.04	0.67±0.03	0.732±0.03
GB	0.79	0.80±0.05	0.67±0.03	**0.733±0.04**
MLP	0.78	0.79±0.03	0.67±0.03	0.729±0.03

(b) PrejudicePT-br. Features: 9, Time: 5.8s - PrejudiceTelegram.br

Method	Auc Score	Precision	Recall	F1-score
LR	0.67	0.72±0.03	0.53±0.04	0.615±0.03
BNB	0.66	0.73±0.04	0.52±0.03	0.612±0.03
MNB	0.65	0.71±0.04	0.52±0.03	0.607±0.03
LSVM	0.67	0.72±0.03	0.53±0.04	0.614±0.03
KNN	0.65	0.65±0.05	0.61±0.08	0.627±0.02
SGD	0.66	0.69±0.06	0.54±0.09	0.598±0.02
RF	0.70	0.71±0.04	0.58±0.02	**0.640±0.02**
GB	0.70	0.71±0.04	0.56±0.04	0.630±0.03
MLP	0.67	0.68±0.04	0.53±0.05	0.597±0.02

Table 9: Results from Using the 842 Words of the PrejudicePT-br Dictionary.

(a) PrejudicePT-br. Features: 842, Time: 34.71s - PrejudiceWhatsApp.Br.

Method	Auc Score	Precision	Recall	F1-score
LR	0.90	0.92±0.00	0.71±0.04	0.804±0.02
BNB	0.90	0.89±0.00	0.74±0.03	0.809±0.02
MNB	0.89	0.88±0.00	0.75±0.04	0.812±0.02
LSVM	0.89	0.91±0.02	0.72±0.04	**0.812±0.03**
KNN	0.83	0.84±0.10	0.70±0.09	0.753±0.01
RF	0.89	0.90±0.01	0.72±0.04	0.802±0.03
SGD	0.89	0.94±0.01	0.69±0.05	0.797±0.03
MLP	0.89	0.91±0.00	0.72±0.04	0.809±0.02

(b) PrejudicePT-br. Features: 842, Time: 24.8s - PrejudiceTelegram.br.

Method	Auc Score	Precision	Recall	F1-score
LR	0.80	0.83±0.03	0.60±0.02	0.701±0.02
BNB	0.79	0.80±0.03	0.63±0.02	0.709±0.02
MNB	0.80	0.80±0.03	0.64±0.02	**0.713±0.02**
LSVM	0.80	0.82±0.02	0.61±0.02	0.703±0.02
KNN	0.74	0.74±0.09	0.64±0.07	0.680±0.01
RF	0.78	0.80±0.02	0.61±0.02	0.695±0.01
SGD	0.79	0.86±0.03	0.56±0.01	0.682±0.01
MLP	0.79	0.81±0.01	0.63±0.01	0.713±0.01

work, we intend to compare the proposed approach with state-of-the-art Large Language Models, such as GPT-4 and DeepSeek-V3.

ACKNOWLEDGMENTS

This work was partially funded by Lenovo as part of its R&D investment under the Information Technology Law. The authors would like to thank LSBD/UFC for the partial funding of this research.

REFERENCES

Bahgat, M., Wilson, S., and Magdy, W. (2022). Liwc-ud: Classifying online slang terms into liwc categories. In *Proceedings of the 14th ACM Web Science Conference 2022*, WebSci '22, page 422–432, New York, NY, USA. Association for Computing Machinery.

Carvalho, F., Junior, F. P., Ogasawara, E., Ferrari, L., and Guedes, G. (2023). Evaluation of the brazilian portuguese version of linguistic inquiry and word count 2015 (bp-liwc2015). *Lang. Resour. Eval.*, 58(1):203–222.

Carvalho, F., Santos, G., and Guedes, G. P. (2018). Affectpt-br: an affective lexicon based on liwc 2015. In *2018 37th International Conference of the Chilean Computer Science Society (SCCC)*, pages 1–5.

CROCHÍK, J. L. (1997). *Preconceito, Indivíduo e Cultura*. Robe Editorial, São Paulo, 1 edition.

de Sá, I. C., Gadelha, T., Vinuto, T., da Silva, J. W. F., Monteiro, J. M., and Machado, J. C. (2023a). A real-time platform to monitoring misinformation on telegram. In Filipe, J., Smialek, M., Brodsky, A., and Hammoudi, S., editors, *Proceedings of the 25th International Conference on Enterprise Information Systems, ICEIS 2023, Volume 1, Prague, Czech Republic, April 24-26, 2023*, pages 271–278. SCITEPRESS.

de Sá, I. C., Galic, L., Franco, W., Gadelha, T., Monteiro, J. M., and Machado, J. C. (2023b). BATMAN: A big data platform for misinformation monitoring. In Filipe, J., Smialek, M., Brodsky, A., and Hammoudi, S., editors, *Proceedings of the 25th International Conference on Enterprise Information Systems, ICEIS 2023, Volume 1, Prague, Czech Republic, April 24-26, 2023*, pages 237–246. SCITEPRESS.

Dinu, L. P., Iordache, I.-B., Uban, A. S., and Zampieri, M. (2021). A computational exploration of pejorative language in social media. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3493–3498, Punta Cana, Dominican Republic. Association for Computational Linguistics.

GOLDSTEIN, J. (1983). *Psicologia social*. Guanabara, Rio de Janeiro, 1 edition.

Leite, J. A., Silva, D., Bontcheva, K., and Scarton, C. (2020). Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In Wong, K.-F., Knight, K., and Wu, H., editors, *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924, Suzhou, China. Association for Computational Linguistics.

MEZAN, R. (1998). *Tempo de muda: ensaios de psicanálise*. Cia das Letras, São Paulo, 1 edition.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pranckevičius, T. and Marcinkevičius, V. (2017). Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Baltic Journal of Modern Computing*, 5(2):221.

Rosenfeld, A., Sina, S., Sarne, D., Avidov, O., and Kraus, S. (2018). A study of whatsapp usage patterns and prediction models without message content. *arXiv preprint arXiv:1802.03393*.

Rubin, V. L., Chen, Y., and Conroy, N. K. (2015). Deception detection for news: three types of fakes. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.

Sert, B. and Ülker, S. V. (2023). A review of liwc and machine learning approaches on mental health diagnosis. *Social Review of Technology and Change*, 1(2):71–92.

Taso, F., Reis, V., and Martinez, F. (2023). Sexismo no brasil: análise de um word embedding por meio de testes baseados em associação implícita. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 53–62, Porto Alegre, RS, Brasil. SBC.

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., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7174–7183, Marseille, France. European Language Resources Association.

Vargas, F., Rodrigues de Góes, F., Carvalho, I., Benevenuto, F., and Pardo, T. (2021). Contextual-lexicon approach for abusive language detection. In Mitkov, R. and Angelova, G., editors, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1438–1447, Held Online. INCOMA Ltd.

Widmann, T. and Wich, M. (2023). Creating and comparing dictionary, word embedding, and transformer-based models to measure discrete emotions in german political text. *Political Analysis*, 31(4):626–641.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., and Wesslén, A. (2012). *Experimentation in software engineering*. Springer Science & Business Media.

Zhu, J., Zhang, Z., Guo, Z., and Li, Z. (2024). Sentiment classification of anxiety-related texts in social media via fuzing linguistic and semantic features. *IEEE Transactions on Computational Social Systems*, pages 1–11.