






Natural Language Processing Techniques for Hate Speech Evaluation for Brazilian Portuguese

Cássia C. S. Rosa^(✉) , Fábio V. Martinez^{}, and Renato Ishii^{}

Faculdade de Computação, Universidade Federal de Mato Grosso do Sul,
Campo Grande, Brazil
{cassia.c,fabio.martinez,renato.ishii}@ufms.br

Abstract. The numerous harmful publications generated from large amounts of data expelled daily on social media make it necessary to adopt automated technologies for online content moderation. Sentence classification and sentiment analysis are Natural Language Processing (NLP) techniques used to detect hate speech on social media platforms such as Facebook and Instagram. However, some difficulties reduce the effectiveness of these tools in the Portuguese language. Previous research has shown how NLP models have high accuracy when trained with datasets centered on mastering the Brazilian Portuguese language. In this work, we propose the creation of a large-scale linguistic corpus for Brazilian Portuguese composed of publications collected from the social network Twitter. The experiments were performed by tuning a pretrained transformer model.

Keywords: Hate Speech · Brazilian Portuguese · NLP

1 Introduction

From the creation of social networks, assumptions about the formation of these spaces began to be studied. Among them, the idealization that a social network would be an environment free of social problems such as racism, misogyny, and LGBTQIA+Phobia stands out. However, it is possible to see that this is a flawed estimate, most likely because the Internet is not fully inclusive, as according to the UN, 37% of the world's population does not have access to the World Wide Web. Thus, it is notable that there is a predominance of only a fraction of society on the Internet since its construction and, as this is a hegemonic space, it is noticeable the engendering of discrimination not only through social networks, but also through blogs, websites, and forums. See for instance the case of the forum called “StormFront” [7], created in 1995 and removed from the Internet in 2017 for propagating Nazi and white supremacist ideas.

Removing content from the Internet is not an easy task, as there are a large number of devices connected daily and countless posts made by them. For this and other reasons, social networks began to use mechanisms capable of automatically detecting violations of their terms of use, causing failures where these

systems ignored complaints, classified them incorrectly, or even took a long time to remove some discriminatory content. In this way, as the use of social networks and other online socialization platforms increases, the need for more efficient and fair systems for detecting and removing offensive content is growing.

A strategy often applied for this purpose is the use of Natural Language Processing (NLP), a branch of the Artificial Intelligence (AI) area that aims to understand human languages automatically. Several techniques such as Bag of Words, Naïve Bayes, and Support Vector Machines (SVM) have been created and are now used for sentiment analysis, construction of search tools, and automatic correction of words and texts. Among the extensive purposes of the NLP, the detection of hate speech is still being consolidated due to the difficulty in detecting the presence of hate in a sentence. This is due both to the lack of consensus on the concept of hate speech and also to the lack of balanced databases that contain different types of hate speech correctly cataloged.

This paper is organized as follows. Section 2 displays important background for the foundation of the research with highlights of some work carried out previously and some problems faced through the analysis of hate speech. Section 3 shows the methodology applied for data collection and annotation. Finally, in Sect. 4 the results are presented, and Sect. 5 the conclusions and future work are discussed.

2 Preliminaries

In this section, important concepts for developing this work will be presented, as well as the techniques and technologies used and their characteristics.

2.1 Natural Language Processing

Natural Language Processing (NLP) is a branch of Artificial Intelligence that aims to understand human languages in an automated way. The Bag of Words [18] is a technique that counts the frequency of words in a text and rearranges them into a list in which information about the word order is discarded. Thus, it is possible to use this result as an input to machine learning models like Naïve Bayes [29], which has a probability table containing the ratio between the occurrence of words and their total number in each corpus. Support Vector Machines (SVM) [10] is another machine learning model that divides the data space into dimensions according to the number of classes of the problem to be treated. The objective of this model is to maximize the margin between the generated spaces to correctly categorize as much data as possible. Other NLP techniques such as lemmatization and stemming are adopted during the data preprocessing stage and are used to reduce and normalize the large volume of words that will be analyzed in the future. While lemmatization transforms a word into its base, stemming transforms a word into its trunk or root.

Despite being widely used in other applications, the techniques mentioned above may be ineffective for detecting hate speech. Because of figures of speech

such as ambiguity, sarcasm, and irony, the context of a sentence is essential in determining toxicity. Therefore, there is an urgent need to use more sophisticated techniques for evaluating hate speech, such as those mentioned below.

A *transformer* [28] is an artificial neural network architecture that has an encoding and decoding framework. The Bidirectional Encoder Representations of Transformers (BERT) [14] is a model based on transformer architecture that has achieved state-of-the-art and was developed by Google’s Artificial Intelligence team to be applied in NLP tasks. It uses masked language modeling (MLM), where some vocabulary words are encoded and the model tries to predict them from their context. Normally, MLM uses unidirectional models, in which the context is understood from left to right or from right to left. BERT, however, uses a bidirectional approach, which combines unidirectional contexts.

Tokenizing a string is an important process for linguistic models (LM). In particular, BERT performs the subtokenization of word pieces, a process in which a word is divided into its prefixes, if they exist, and, with the result of this segmentation, the identified terms are called *tokens*. Hence, the generated tokens are provided as input to the model that performs the *embedding activity*, or *embedding*, to represent them as a vector of weights in which each position represents information about them. The tokenization step is assigned as input to the LM which has 12 layers in its basic form and 24 layers in its extensive form.

BERTimbau [26] is a BERT model trained for Brazilian Portuguese that initially implements text similarity, text element recognition, and named entity recognition tasks. In this work, this model was adapted to classify sentences that contain some type of hate speech.

2.2 Hate Speech

The Brazilian Law Number 7716, of January 5, 1989, defines that acts of discrimination or prejudice based on race, color, ethnicity, religion, or national origin resulting from hate speech must be criminalized. However, the crimes of racism and racial/discriminatory injury are seen as different because, in the first, it is understood that the aggressor’s intention would be to offend the entire group to which the victim belongs, while in the second, it is understood that his intention would be to offend only the victim. As in the legal sphere, these definitions contribute to the emergence of problems that can make it difficult to classify hate speech.

Thus, to accomplish the detection of hate speech, it is necessary to have a well-defined parameter about what are the main characteristics of a manifestation of hate and how it can manifest itself in different situations. Hate speech can be identified through two basic characteristics, which are the insult or offense to a person who is included in a socially vulnerable group and speeches, gestures, or expressions that incite violence explicit or implicit [22].

Another ingredient that also contributes to wrong classifications of hate speech is the way in which the aggressor manages to disguise the hatred through sentences that can be considered opinions, political ideologies, or freedom of speech. Rosenfield [23] defines two ways of expressing hate. For him, *hate speech*

in form are explicitly hateful manifestations, such as racist insults aimed at racial groups. While *hate speech in substance* are veiled manifestations of hatred, such as denial of the Holocaust or other type of message that does not contain injuries.

Because there are still many discussions on the subject, in this work only explicitly hateful manifestations will be classified as hate speech, since this is a generally well-defined category in research that surround this subject.

2.3 Related Works

The debate on hate speech is covered by several areas of knowledge such as Computer Science, Law, and Literature. In the former, several researchers seek to improve their techniques for evaluating words, phrases, and speeches to contribute to less toxic virtual environments. However, the majority of research is developed by English-based texts [6, 12, 13, 30], making it difficult for these systems to operate in other languages [1, 3, 8, 9]. In this way, to increase the performance of hate speech detection in Portuguese, some previous works were done.

Pelle and Moreira [21] collected 10,000 comments from the Brazilian news page called G1 and created two databases: OFFCOMBR-2, in which at least two annotators agreed with the chosen class, and OFFCOMBR -3, where all three annotators agreed with the chosen class, namely “offensive” and “non-offensive”. The “offensive” class is categorized into the subclasses “racism”, “sexism”, “xenophobia”, “homophobia”, “religious intolerance”, and “swearing”. The Naïve Bayes and SVM algorithms were used to classify these comments. The best classifier, SVM, achieved a score of 80% on the F measure.

Fortuna and colleagues [16] collected 5,668 tweets to compose their database, which was divided into “hate speech” and “non-hate speech” classes. The class “hate speech” was subdivided into subtypes called “sexism”, “corporal”, “origin”, “homophobia”, “racism”, “ideology”, “religion”, “health”, and “other lifestyle”. For classification, the LSTMs deep learning model was used, which obtained a score of 78% on the F measure.

Leite et al. [20] presented TOLD-Br, a database with around 21,000 tweets in Brazilian Portuguese. The “toxic” and “non-toxic” classes were used for binary classification, while the “LGBTQ+Phobia”, “xenophobia”, “racism”, “misogyny”, “obscene”, and “insulting” were used for multi-classification. Pretrained natural language processing models together with a model created from Bag-of-Words techniques and automated machine learning were used for both classifications. Their best models, which used the BR-BERT and M-BERT-BR architecture, scored on the F measure of 76% and 75%, respectively.

Silva [25] used the dataset provided by Fortuna et al. In addition, the SVM, MLP Neural Network, Logistic Regression, and Naïve Bayes algorithms had some configurations tested to obtain better performance. The SVM algorithm outperformed the others and obtained an F measure of 71%.

This work arises with the premise of collecting Twitter publications to compose an extensive database, which will be cataloged in three different classes

and will later be applied to detect sentences constituted by some type of explicitly hateful manifestation. A deep learning model will be used to classify the previously collected items.

2.4 Scope

For identifying possible hate speech, it is first necessary to verify whether this is content that discriminates against a person or a minority social group to which he belongs. Even more so if this is content that instigates violence against that person or against socially vulnerable groups. In addition, the dynamics between users, the social, cultural, and regional contexts, as well as the context of the phrase and issues related to the resignification of words among groups marginalized by society, among other issues, need to be considered to guarantee the legitimacy of possible classifications made by linguistic models.

Other issues related to the automation of publication analysis also need to be considered. For Duarte et al. [15], the tools used for moderating content online have certain limitations related to the NLP, such as the need for a specific domain, relatively low reliability and inferior classification ability to humans. Those issues that can disproportionately marginalize and censor discriminated groups, such as with the [Perspective API](#), which ranked Twitter profiles of drag queens as more toxic than profiles of white supremacists and far-right leaders [2].

Algorithmic biases are also present in natural language processing tools and large models [5] used for content moderation on social media platforms. Therefore, to prevent these technologies from being inefficient and biased, it is essential to create a robust, balanced, and correctly annotated database, as well as to train more accurate models and investigate the presence of bias in existing datasets [11, 19]. Together, the results obtained by these models must be evaluated by humans to guarantee their correctness. Furthermore, it is important to emphasize that discrimination also happens outside of social networks and the application of hate detectors is just a small step towards mitigating discrimination, which needs to be fought in a significant way outside of the Internet.

Because of the amount of data used in the training process, BERT is classified as a Large Language Model (LLM). The high consumption of resources like energy to train these models can cause big environmental impacts [27]. Furthermore, ethical and social issues can be effect by biased dissemination [31] arising from a lack of representativeness on data [5]. Additionally, the identity terms used in this work can increase the discrimination and marginalization of minority groups [11]. This occurs because offensive words and expressions could have different meanings, depending on the context in which they are applied [17]. Thus, it is important to invest in mitigation and evaluation of bias techniques in language models and datasets [32].

3 Databases

In this section, we described how the construction of the database was accomplished, its annotations, and other characteristics.

3.1 Data Collection and Preprocessing

The *snsrape*¹ is tool for data mining that support social media services like Facebook, Instagram, Reddit, and Telegram. Differently from the *Twitter API*, available only to developers, this is an open-source tool that can be used by anyone and does not require registration, as well as there is no need to perform any type of authentication beforehand and there is also no limit for requests, nor a maximum capacity for extracting publications. For these reasons, *snsrape* was the tool chosen to achieve the data collection process.

As the intention is to collect only *hate in form* publications, queries were carried out to capture publications that contain any of the terms mentioned below, which are arguments for a search.

The hashtag “eleições2022” was consulted after the second round of the 2022 Brazilian presidential elections. It was chosen because of its overlap with other hashtags such as “vaidatpt” and “brasilvota22”, which were frequently used in that same period. Publications that contained the terms “favelado”, “nordestino”, “indigenous” or “african” were consulted in the months of October and November, with the intention of capturing possibly xenophobic terms. In this same time interval, posts consisting of the terms “vagabunda” (slut), “piranha” (whore), “feminazi” or “macumbeira” (witch) were also consulted. The words “humor negro” (black humor), “negro” and “escravo” (slave) were searched between August and November to evaluate publications with racist content. Adjectives commonly used in ableist sentences, such as “retardado” (retarded), “débil” (moron) and “demente” (demented), were analyzed from September to November of the same year. As well as “obeso” (obese) and “gordo” (fat), which were also surveyed in the same period. Finally, more broadly, the terms “traveco” (tranny), “afeminado” (effeminate), and “opção sexual” (sexual option) were used to capture tweets from January 2019 to November 2022. The step-by-step of this process can be seen in Fig. 1. Also, the word cloud of the entire corpus is show in Fig. 2.

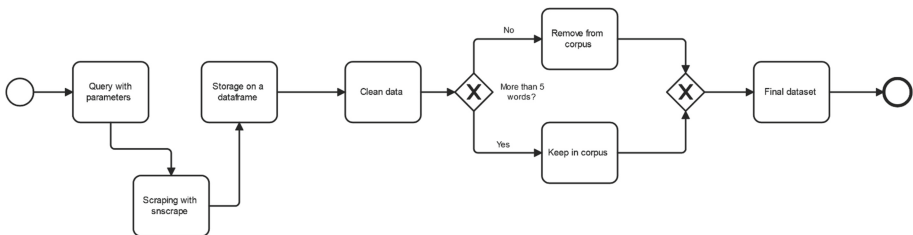


Fig. 1. Process of data collection and data cleaning.

¹ <https://github.com/JustAnotherArchivist/snsrape>.

number of words mistakenly annotated as neutral, which may trigger a possibility that implicitly hateful speeches be treated as non-toxic. Furthermore, tweets classified as neutral were not necessarily identified as positive, as they may contain offensive words, but which are not intended for a person or are not harmful in the context used.

To clarify the classification performed in this work, Table 1 shows some annotated examples.

Table 1. Examples of annotated sentences in Portuguese. Three types of annotations are presented: Hate speech, Offense, and Neutral.

Text in Portuguese (English translation)	Annotation
Que negros de merda (What fucking niggers)	Hate speech
Sua loira vagabunda do caralho morte p vc é pouco... (You fucking blonde slut death to you is little...)	Hate speech
Essa enquete prova que o brasileiro é um povo MT retardado... (This poll proves that Brazilians are a VERY retarded people...)	Offense
tenho muito nojo de gordo suado cheio de banha kkk (I'm really disgusted with sweaty fat people full of lard lol)	Offense
Amiga mas nem todo mundo tem o sonho de ser piranha (Friend, but not everyone has the dream of being a slut)	Neutral
Homens negros falem sobre amor, isso motiva outros negros (Black men talk about love, it motivates other black people)	Neutral

Furthermore, Fig. 3 shows the process of data annotation adopted in this work.

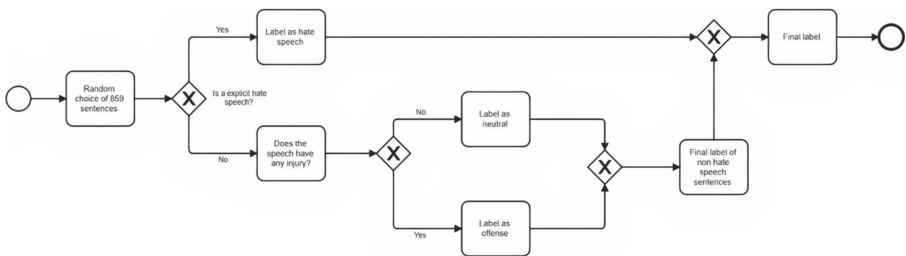


Fig. 3. Process of data annotation.

An exploratory text data analysis was conducted in the annotated dataset to acquire more knowledge about it. Figure 4 shows the most frequent words present on the dataset. By this, it is possible to recognize that the dispersion of data is compatible with the chosen words and hashtags used during the collection

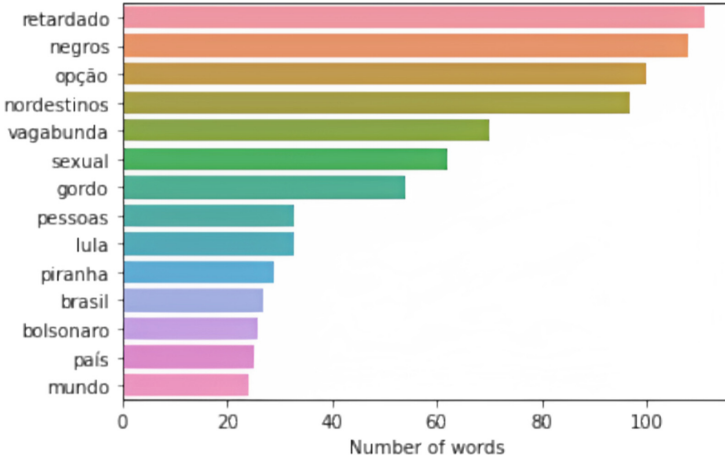


Fig. 4. The most frequent words in the annotated dataset.

process. Thus, the sentences selected to compose the annotated dataset represent the original corpus.

To exemplify the composition of data in each category, i.e. neutral, offensive, and hateful, four bigrams were created. In Fig. 5, except for “sexual option”, which predominates in all sets, the other bigrams demonstrate, mainly, the differences between the definitions of hate speech and offense. Notice that, in the hate speech class, there is a high correlation between the bigrams and misogyny speech.

As seen in Fig. 6, the variation between the number of words present in a sentence in the annotated data set is also representative of the evaluation. This is because the tweets produced, containing hate or not, do not follow a pattern and can vary in different ways, one of which is the number of words. However, it is also possible to notice that as the number of words increases, the number of tweets decreases, since social media users usually post with less than 30 words.

4 Results

Due to the small number of labeled publications, the *hatecheck-portuguese* dataset, by Röttger and colleagues [24], which contains 2581 sentences annotated as *hateful* and 1110 annotated as *not-hateful*, was chosen for fine-tuning the pretrained model BERTimbau. This is the process in which a deep learning model has its parameters adjusted to perform tasks from a different dataset than the one it was previously trained on. Data were originally used for functional tests obtained in the language model for sentiment analysis XLM-T [4] and Google’s Perspective API. Functional testing is an evaluation performed to measure the

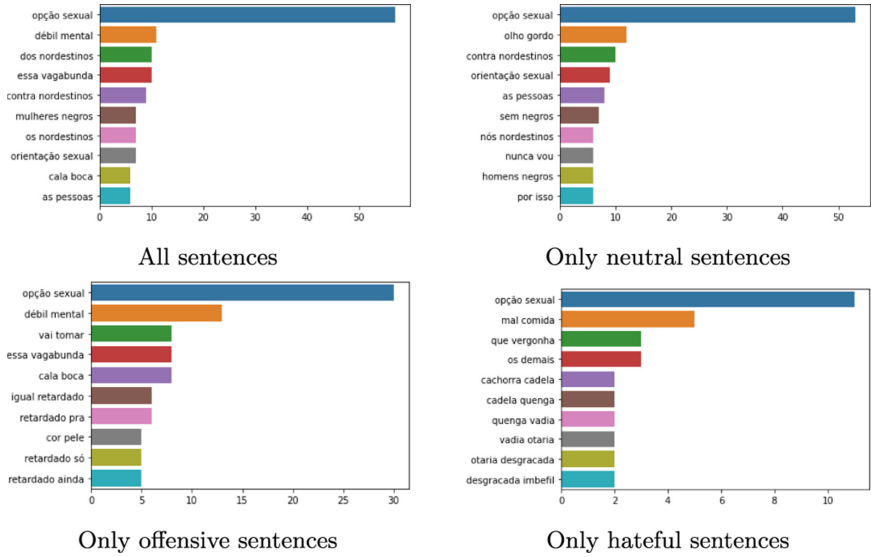


Fig. 5. The composition of data as a bigram in each category of annotated dataset: Neutral, Offensive, and Hateful.

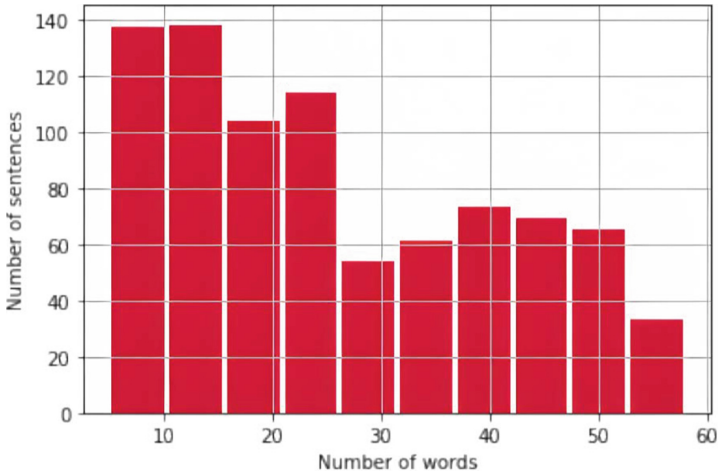


Fig. 6. Number of words per sentence in the annotated dataset.

quality of software through the analysis of the outputs generated by a program from a certain set of data provided as input.

Determining factors for choosing this database were (1) the examples contained in the corpus “hateful” being explicitly hateful manifestations, and (2) the process of generating test cases, done manually by specialists in the Portuguese language who have experienced previous research or annotation of hate

speech. Additionally, the OffComBR-2 [21] database, which contains a corpus of 831 examples annotated as ‘no’, and a corpus of 419 examples annotated as ‘yes’, was also used to compose the evaluation step. Although the functional test set has more examples, the texts extracted from news comments by Pelle and Moreira [21] have a structure similar to that of the collected tweets and, for this reason, can positively influence the learning of the model.

In this way, the BERTimbau adjustment was performed from these data, and each set was applied separately to generate a model. In both sets, a random division of the data of 80% for training and 20% for validation was performed. Fine-tuning was defined in 4 steps with batch sizes equal to 16 and a learning rate starting at 0.00001. The resulting models were tested with collected data that were labeled “neutral” and “hate speech”. As seen in Fig. 7, the first model using the functional test set, obtained an F measure of 68% and a accuracy of 66%. The second model, using the set of offensive comments, obtained an F measure of 74% and a accuracy of 73%. Another test was also performed with the second model to evaluate the “offense” corpus of the collected database. The F measure remained at 74% while the accuracy increased to 74%. The F measure, or simply F1, is a metric for evaluating the performance of machine learning models. More specifically, it is responsible for observing other metrics that it evaluates: precision and recall, which are used to measure how well the model performed in avoiding false positives and false negatives, respectively. The Table 2 helps to elucidate the performance of all models and points out how even with few examples, the collected database has a huge potential for hate speech detection task. All results obtained can be found at [TwitterHateBR](#).

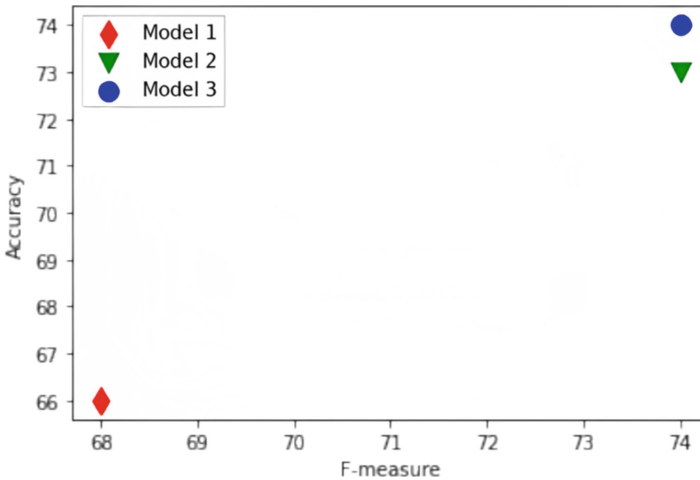


Fig. 7. F-measure and Accuracy of each model.

Table 2. Each model and his respective metrics and database. Two types of metrics are presented: F-measure and accuracy. Three databases are used: Functional test set, offensive comments and the collected database.

Model	F-measure	Accuracy	Database
Model 1	68%	66%	Functional test set
Model 2	74%	73%	Offensive comments
Model 3	74%	74%	Collected database

5 Conclusion

As with other AI problems, to perform hate speech detection, determining a database is an essential step in training good models. Therefore, it is important to collect large amounts of examples belonging to the analyzed domain, as well as to correctly catalog them based on the fundamentals and discussions about expressions of hate.

Like the previous, this work has been faced with obstacles to find examples for the hateful class, since the vast majority of the sentences contained in databases are neutral or offensive. This imbalance, in addition to generating false classifications, can also bring bias, since the low amount of data makes the model not able to learn correctly the hate speech concept.

This work presented the process achieved to collect Twitter publications used to create a large database in Brazilian Portuguese, as well as its criteria for evaluating instances and labeling of sentences called hate speech. The result of applying the database created to two models trained from previous works is also presented, in which the F measure obtained in the second model is within the range found in previous works.

Future work consists of including the annotation of more tweets by volunteers with previous experience in hate speech annotation and the fine-tuning of other pretrained models from this data. Additionally, it is feasible to apply techniques of evaluation and mitigation of bias in the dataset and models and also to accomplish adjustments in the dataset to reach the balance of classes.

Acknowledgments. We thank the support of the UFMS (Universidade Federal de Mato Grosso do Sul). We also thank the support of the INCT of the Future Internet for Smart Cities funded by CNPq, proc. 465446/2014-0, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, and FAPESP, proc. 2014/50937-1 and 2015/24485-9.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of, FAPESP, CAPES, and CNPq.

References

1. Aluru, S.S., Mathew, B., Saha, P., Mukherjee, A.: Deep learning models for multilingual hate speech detection. arXiv preprint [arXiv:2004.06465](https://arxiv.org/abs/2004.06465) (2020)
2. Antonialli, D.: Drag queen vs. David Duke: Whose tweets are more ‘toxic’. Wired. Retrieved (July/August 2019)
3. Plaza-del Arco, F.M., Molina-González, M.D., Urena-López, L.A., Martín-Valdivia, M.T.: Comparing pre-trained language models for Spanish hate speech detection. *Expert Syst. Appl.* **166**, 114120 (2021)
4. Barbieri, F., Anke, L.E., Camacho-Collados, J.: XLM-T: multilingual language models in twitter for sentiment analysis and beyond. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 258–266 (2022)
5. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: can language models be too big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623 (2021)
6. Biere, S., Bhulai, S., Analytics, M.B.: Hate speech detection using natural language processing techniques. Master Business Analytics Department of Mathematics Faculty of Science (2018)
7. Bowman-Grieve, L.: Exploring “stormfront”: a virtual community of the radical right. *Stud. Conflict Terror.* **32**(11), 989–1007 (2009)
8. Corazza, M., Menini, S., Cabrio, E., Tonelli, S., Villata, S.: Cross-platform evaluation for italian hate speech detection. In: *CLiC-it 2019–6th Annual Conference of the Italian Association for Computational Linguistics* (2019)
9. Corazza, M., Menini, S., Cabrio, E., Tonelli, S., Villata, S.: A multilingual evaluation for online hate speech detection. *ACM Trans. Internet Technol.* **20**(2), 1–22 (2020)
10. Cortes, C., Vapnik, V.: Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995)
11. Davidson, T., Bhattacharya, D., Weber, I.: Racial bias in hate speech and abusive language detection datasets. arXiv preprint [arXiv:1905.12516](https://arxiv.org/abs/1905.12516) (2019)
12. Davidson, T., Warmley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, pp. 512–515 (2017)
13. De Gibert, O., Perez, N., García-Pablos, A., Cuadros, M.: Hate speech dataset from a white supremacy forum. arXiv preprint [arXiv:1809.04444](https://arxiv.org/abs/1809.04444) (2018)
14. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805) (2018)
15. Duarte, N., Llanos, E., Loup, A.: Mixed messages? The limits of automated social media content analysis. In: Friedler, S.A., Wilson, C. (eds.) *Proceedings of the 1st Conference on Fairness, Accountability and Transparency. Proceedings of Machine Learning Research*, vol. 81, pp. 106–106. PMLR (23–24 February 2018)
16. Fortuna, P., Rocha da Silva, J., Soler-Company, J., Wanner, L., Nunes, S.: A hierarchically-labeled Portuguese hate speech dataset. In: *Proceedings of the Third Workshop on Abusive Language Online*, pp. 94–104. Association for Computational Linguistics, Florence, Italy (August 2019)
17. Garg, T., Masud, S., Suresh, T., Chakraborty, T.: Handling bias in toxic speech detection: a survey. *ACM Comput. Surv.* (2023, just accepted). <https://doi.org/10.1145/3580494>
18. Harris, Z.S.: Distributional structure. *Word* **10**(2–3), 146–162 (1954)

19. Huang, X., Xing, L., DERNONCOURT, F., Paul, M.J.: Multilingual twitter corpus and baselines for evaluating demographic bias in hate speech recognition. arXiv preprint [arXiv:2002.10361](https://arxiv.org/abs/2002.10361) (2020)
20. Leite, J.A., Silva, D.F., Bontcheva, K., Scarton, C.: Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. arXiv preprint [arXiv:2010.04543](https://arxiv.org/abs/2010.04543) (2020)
21. de Pelle, R., Moreira, V.: Offensive comments in the Brazilian web: a dataset and baseline results. In: Anais do VI Brazilian Workshop on Social Network Analysis and Mining. SBC, Porto Alegre, RS, Brasil (2017)
22. Rocha, J.L.A., Mendes, A.P.T.: Guidance Booklet for Victims of Hate speech (in Portuguese) (2020)
23. Rosenfeld, M.: Hate speech in constitutional jurisprudence: a comparative analysis. Soc. Sci. Res. Netw. **41**, 1–63 (2001)
24. Röttger, P., Seelawi, H., Nozza, D., Talat, Z., Vidgen, B.: Multilingual hateCheck: functional tests for multilingual hate speech detection models. arXiv preprint [arXiv:2206.09917](https://arxiv.org/abs/2206.09917) (2022)
25. Silva, A.S.R.: Study of distributional models for detecting hate speech in Portuguese.. Ph.D. thesis, Universidade de São Paulo (2021). (in Portuguese)
26. Souza, F., Nogueira, R., Lotufo, R.: BERTimbau: pretrained BERT Models for Brazilian Portuguese. In: Cerri, R., Prati, R.C. (eds.) Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I, pp. 403–417. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-61377-8_28
27. Strubell, E., Ganesh, A., McCallum, A.: Energy and policy considerations for deep learning in nlp. arXiv preprint [arXiv:1906.02243](https://arxiv.org/abs/1906.02243) (2019)
28. Vaswani, A., et al.: Attention is all you need. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
29. Vikramkumar, B.V., Trilochan: Bayes and naive bayes classifier. CoRR abs/1404.0933 (2014)
30. Waseem, Z., Hovy, D.: Hateful symbols or hateful people? Predictive features for hate speech detection on twitter. In: Proceedings of the NAACL Student Research Workshop, pp. 88–93 (2016)
31. Weidinger, L., et al.: Ethical and social risks of harm from language models. arXiv preprint [arXiv:2112.04359](https://arxiv.org/abs/2112.04359) (2021)
32. Xia, M., Field, A., Tsvetkov, Y.: Demoting racial bias in hate speech detection. arXiv preprint [arXiv:2005.12246](https://arxiv.org/abs/2005.12246) (2020)