Challenging Linguistic and Cultural Diversity: Evaluation of AI Models in the Detection of Hate Speech in Brazilian Social Networks

Annie Amorim[1], **Gabriel Assis**[1], **Jonnathan Carvalho**[2],
Daniel de Oliveira[1], **Daniela Vianna**[3], **Mariza Ferro**[1] **and Aline Paes**[1]
[1] Institute of Computing, Universidade Federal Fluminense, Niterói, RJ, Brazil
[2] Department of Informatics, Instituto Federal Fluminense, Itaperuna, RJ, Brazil
[3] JusBrasil, Brazil
{*annieamorim,assisgabriel*}*@id.uff.br*, *joncarv@iff.edu.br*,
{*danielcmo,mariza,alinepaes*}*@ic.uff.br*, *dvianna@gmail.com*

1 Introduction

Social media platforms offer a way to spread information at an unprecedented pace, expanding the scope and effectiveness of how we communicate and share opinions (Moura, 2016; Pelle et al., 2018). Nowadays, they act as a platform for public discussion and debate, allowing people to share their perspectives among individuals or groups (Amorim et al., 2022). Their environment fosters the expression of multiple views on a diverse variety of everyday issues (Paiva et al., 2019). However, this is not without drawbacks: enhancing the propagation of offenses and hate speech are some of their negative side (Aluru et al., 2020). In this context, offensive comments are defined as those that contain any communication affronting one or more individuals, ranging from using inappropriate vocabulary to direct insults (Pelle et al., 2018). Hate speech, on the other hand, is characterized as any expression that intends to offend a third party or a group based on characteristics such as ethnicity, race, nationality, sexual orientation, and gender (Paiva et al., 2019; Vargas et al., 2021). Unfortunately, those comments quickly reverberate through social media (Saraiva et al., 2021).

This paper focuses on this issue in Brazil, the largest and most populous country in Latin America, where some cases of hate speech can be framed as a crime[1]. However, detecting hate speech on social media poses a significant challenge due to several factors, including the large volume of data, the complexity imposed by the wide linguistic diversity, and the expressions unique to the context of these platforms (Vargas et al., 2021). This scenario calls for sophisticated approaches to classification, as offered by Artificial Intelligence (AI), particularly Machine Learning models, which demonstrate significant potential by identifying patterns in varied contexts (Paiva et al., 2019). Pre-trained transformer-based models provide a possible efficient solution to these challenges, which can be tuned for specific tasks. Such fine-tuning of the models not only saves time, resources, and energy, but is also crucial considering the substantial size of these models and the intense GPU demand for training from scratch (Jahan and Oussalah, 2023). For the development of this research, fine-tuned classifiers were adopted, based on BERT and BART models. However, several studies raise doubts about the generalizability of these models and about the costs associated with re-training in the face of changes in data (Yin and Zubiaga, 2021). This work is also dedicated to evaluating the generalization capacity of models in different hate speech datasets.

Most previous work has focused on developing methods to detect hate speech in English (Jahan and Oussalah, 2023). It is essential to devote special attention to the study of hate speech in other languages, since methods developed for one language may be inappropriate when applied to others without the necessary adaptations (Souza et al., 2020a). This is due to the particularities, nuances, and unique expressions that characterize the language. This paper addresses the Portuguese language, the official language spoken in Brazil. Brazil's cultural diversity, influenced by experiences, culture, traditions, and history of colonization, presents additional challenges. This linguistic and cultural diversity is crucial in the ethical analysis of models, questioning whether a model, even one trained explicitly for the Portuguese and contextualized to Brazil, can adequately encompass the country's wide range of linguistic and cultural variations. This work aims to analyze the efficiency of AI models in a ternary setting of hate speech classification, focusing on generalization to new data, getting closer to the real context. In this way, comments are classified as neutral, offensive, or hate speech. It focuses particularly on the qualitative analysis of the results to try to understand how the

[1] https://bit.ly/planalto-lei-7716

models handle such phenomena. Specifically, it seeks to understand the ethical implications of the results, especially how errors can unduly censor a publication or fail to protect vulnerable groups. Although some digital platforms have their own prevention systems, they have limitations (Yin and Zubiaga, 2021; Assis et al., 2024).

Disclaimer: this paper includes offensive and toxic texts as examples of the problems discussed.

2 Model Selection

BERT and BART-based models are tuned for classification tasks using the fine-tuning strategy (Souza et al., 2020b; Lewis et al., 2019). The number of epochs is limited to 2, to save computational resources, complemented by a weight decay of 0.01. Although the number of epochs is chosen to reduce costs, the other parameters follow the default values. Among the models, four are variants of the BERT, adapted and trained specifically with corpora in Portuguese of Brazil: (i) BERTimbau (Souza et al., 2020b), (ii) AlBERTina PT BR (Rodrigues et al., 2023), (iii) BERTweet.BR (Carneiro, 2023) and (iv) DistilBERT PT[2]. Both BERTimbau and AlBERTina were trained with texts of a more formal nature, while BERTweet.BR used a corpus of tweets. The BERTimbau model is tried in two variants, 'base' and 'large'. DistilBERT PT is an optimized version of BERTimbau, obtained through a knowledge distillation process. In addition, two multilingual models are evaluated: (v) BERT Multilingual (mBERT) (Devlin et al., 2018) and (vi) DistilBERT Multilingual (Sanh et al., 2019), the first of which was pre-trained with texts in 104 languages from Wikipedia and the second is a condensed version of BERT Multilingual. Finally, (vii) BART PT was pre-trained with texts in Portuguese.

3 Experimental Evaluation

The classes were numerically coded: 'neutral' as 0, 'offensive' as 1, and 'hate speech' as 2. Data preparation involves eliminating duplicate text and replacing user mentions with the term @USER, links with HTTPURL, and emojis with their respective textual representations.

3.1 Quantitative Assessment

The dataset used for this evaluation was HateBR (Vargas et al., 2022), consisting of comments on Instagram. The assessment adopts stratified cross-validation with $k = 10$. The results presented in Table 1 highlight that the BERTweet.BR and BERTimbau Large models achieved the highest values of F1 for class 2 and F1 Macro, suggesting a better classification performance. It is noted that BERT-based models trained specifically for Portuguese performed better compared to multilingual models. However, BART PT, although trained in Portuguese, could not outperform multilingual models. Both DistilBERT Multilingual and BART showed the lowest results.

Table 1: Results of the models in the HateBR set.

Model	Accuracy	F1 Class 2	F1 Macro	SD F1 Macro
BERTimbau Base	0.858	0.708	0.819	0,024
BERTimbau Large	0.874	0.761	0.844	0,020
BERTweet.BR	0.885	0.746	**0.850**	0,025
AlBERTina BR	0.833	0.686	0.782	**0,179**
DistilBERT PT	0.843	0.666	0.798	0,021
DistilBERT Mult	0.793	0.575	0.736	0,020
BERT Mult	0.822	0.660	0.780	0,038
BART PT	0.814	0.575	0.752	0,018

3.2 Generalization Assessment

Testing the ability to generalize, precisely by observing errors in data not previously seen, allows for an analysis closer to the real context. The best-performing model from the previous step (BERTweet.BR which had the best F1 Macro - Table 1) is trained using the entire HateBR dataset. Subsequently, the trained model is tested with a different dataset, the ToLD-Br (Leite et al., 2020), achieving an F1 Macro of 0.488. In 35 texts classified as 0, the model identified it as 2, indicating potential annotation errors. These texts included obscene words, offensive themes, or pejorative terms, as shown in Table 2, highlighting the difficulty of classifying social media content, especially when it involves ambiguous or sensitive language. In this context, the distinction between neutral content and hate speech is significant, raising ethical concerns about the possibility of failing to protect vulnerable groups. In addition, 89 texts from class 1 were classified as class 2 by the model, underlining the difficulty in differentiating offensive comments from hate speech, which also manifests itself through offenses. Analyses indicate the presence of sexism in these texts, with examples in Table 2. Texts 7 and 13 may also contain prejudices related to age and weight, respectively.

In 156 cases, texts originally labeled as 2 were rated by the model as 0. Many of these texts included regional expressions or terms that, in certain contexts, may be offensive, but in these specific

[2] https://bit.ly/distilbert-portuguese-cased

cases, they were employed as forms of interjections. In text 15, the term '*macaca*', preceded by '*te amo*', seems to indicate affection. Still, the history and everyday use of the term is often associated with racism, leading labelers to classify it as hate speech. In texts 17 and 21, the term '*viado*', although potentially homophobic, is used as a vocative, common in some regions of Brazil, suggesting a non-offensive context. Examples include congratulations and expressions of support. In addition to texts 18 and 20 with the terms '*pretinha*' and '*neguinha*', which are used together with the possessive pronoun '*minha*', they are used affectionately and intimately. This usage indicates an expression of affection or closeness, showing how words can acquire different connotations depending on the context and the relationship between the people involved. The labeling of texts in this domain is inherently complex, as the definition of hate speech requires consideration of the context, intent, and relationship between those involved. There is a fine line between protecting from potentially offensive language and suppressing cultural expressions that, while they may be interpreted as offensive outside of their specific cultural context, are harmless or even endearing within it. Misclassifying these expressions as hate speech can have implications for freedom of expression and cultural preservation.

Table 2: ToLD-Br texts classified by BERTweet.BR.

ID	Text	Label	Model
1	na internet amiga das trans, na vida real junta com hetero pra falar "traveco".	0	2
2	volteeei pras puta 🤬	0	2
3	@USER @USER viado com certeza...	0	2
4	@USER branco mais fodido do tt :coração_vermelho:	0	2
5	sdd que quando eu era puta, eu era mt mais feliz shishdishdish	0	2
6	rt @USER ae bolsogado. e essa aqui, agora? 🤡 aplaudiram hoje? HTTPURL	0	2
7	@USER @USER @USER sonha mais tiazona.	1	2
8	rt @USER @USER o meu ficou como piranha testuda kkkkkkkkkkkkkk	1	2
9	amo essa cabeçuda tomanno cu meu mubdo todinho.	1	2
10	@USER vcs esquerdeopatas não se emendam.	1	2
11	tô vendo a hora eu tacar fogo nessas burguesa do caralho	1	2
12	rt @USER que piranha gostosa🤤🤤 HTTPURL	1	2
13	@USER anta gorda HTTPURL	1	2
14	rt @USER fofa safada e retardada	1	2
15	@USER grt tu é forte pra crc, desanima não. haverá dias ruins e haverá dias bons! tô aqui, se quiser conversar me chama. te amo macaca :rosto_sorridente_com_olhos_de_coração:	2	0
16	esse aí e carioca kkkkk	2	0
17	rt @USER @USER parabéns viado tmj	2	0
18	voltei a falar com minha neguinha :coração_vermelho:	2	0
19	@USER bicha, é sério? hhahahahahaha	2	0
20	hoje tem a festinha da minha pretinha :corações_girando:	2	0
21	sim viado to precisando	2	0
22	mereço minha irmã falando que virou viado kk	2	1
23	ah coe mano, trair a mulher com um viado??? coe coe	2	1
24	viadinho branco fa de outlander criticando series negras talentosas que nao foram feitas pra ele HTTPURL	2	2
25	@USER bambi kk	2	2
26	rt @USER pelo menos a sapatão não está batendo na outra igual alguns heteros que bate na mulher HTTPURL	2	2

As for the 98 texts labeled as 2 and classified by the model as 1, using terms similar to those of the previous case was observed, but with an offensive intent. Although the model did not correctly classify these cases, it was notable that it could differentiate the categories through their classification. The examples of texts 18 and 21, contrasting with texts 22 and 23, exemplify how the same term '*viado*' can alternate between a use as a vocative and a pejorative, homophobic use. This variation highlights the complexity of discerning contexts and intentions in identifying offensive or hate speech. This example underlines the complexity and challenges in automatic language classification, especially in texts loaded with cultural and contextual nuances. The model correctly identified 36 class 2 texts. Among these, sentence 26 stood out with its use of the term '*bambi*'. While it may be offensive in certain contexts, it also refers to a children's movie. This points to the need for models that understand context to avoid misclassifications, especially in terms with multiple meanings. Models should be trained with diverse data and labeled correctly, to avoid biases that could lead to discrimination or unfairness, while recognizing the diversity of contexts and linguistic uses.

4 Conclusions

BERT models in Portuguese, especially BERTweet.BR for short texts, had accurate results. However, when testing on ToLD-Br, performance dropped, particularly in detecting hate speech, pointing to limitations in generalization. This is due, in part, to the complexity of labeling in cases of hate speech, where regionalisms and interjections can be misinterpreted. In addition, there were problems in identifying sexist speech, which was often classified as neutral by annotators. These findings underline the need for contextual analysis in the evaluation of texts. The misclassification of offensive content and hate speech has significant ethical implications and, therefore, demands prudence and attention. Datasets may have cultural and historical biases that do not encompass linguistic diversity. Brazil, with its rich cultural diversity, illustrates how the perceived offensiveness of terms can vary internally and even more so between countries that share the same language. In addition, automated moderation faces challenges such as false positives, which can lead to undue censorship, and false negatives, which fail to protect vulnerable groups or enforce the law correctly. Therefore, it is essential to recognize that AI should be a tool for supporting, not substituting, content moderation. The accuracy of the identification is vital, given the severe ethical consequences of misclassification.

Acknowledgements

This research was financed by CNPq (National Council for Scientific and Technological Development), grant 307088/2023-5, FAPERJ - *Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro*, process SEI-260003/000614/2023 and SEI-260003/002930/2024, and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

References

Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. Deep learning models for multilingual hate speech detection. *CoRR*, abs/2004.06465.

Annie Amorim, Nils Murrugarra-Llerena, Vítor Silva, Daniel de Oliveira, and Aline Paes. 2022. Modelagem de tópicos em textos curtos: uma avaliação experimental. In *Anais do XXXVII Simpósio Brasileiro de Bancos de Dados*, pages 254–266, Porto Alegre, RS, Brasil. SBC.

Gabriel Assis, Annie Amorim, Jonnatahn Carvalho, Daniel de Oliveira, Daniela Vianna, and Aline Paes. 2024. Exploring Portuguese hate speech detection in low-resource settings: Lightly tuning encoder models or in-context learning of large models? In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 301–311, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Fernando Pereira Carneiro. 2023. Bertweet.br: A pretrained language model for tweets in portuguese. Dissertação de mestrado, Universidade Federal Fluminense, Programa de Pós-Graduação em Computação, Niterói.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546:126232.

João Augusto Leite, Diego Silva, Kalina Bontcheva, and Carolina Scarton. 2020. Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In *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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Marco Aurelio Moura. 2016. *O discurso do ódio em redes sociais*. Lura Editorial (Lura Editoração Eletrônica LTDA-ME).

Peter Paiva, Vanecy da Silva, and Raimundo Moura. 2019. Detecção automática de discurso de ódio em comentários online. In *Anais da VII Escola Regional de Computação Aplicada à Saúde*, pages 157–162, Porto Alegre, RS, Brasil. SBC.

Rogers Pelle, Cleber Alcântara, and Viviane P. Moreira. 2018. A classifier ensemble for offensive text detection. In *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web, WebMedia 2018, Salvador-BA, Brazil, October 16-19, 2018*, pages 237–243. ACM.

João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Osório. 2023. Advancing neural encoding of portuguese with transformer albertina pt-*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Ghivvago Damas Saraiva, Rafael T. Anchiêta, Francisco Assis Ricarte Neto, and Raimundo Santos Moura. 2021. A semi-supervised approach to detect toxic comments. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), Held Online, 1-3September, 2021*, pages 1261–1267. INCOMA Ltd.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020a. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020b. Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.

Francielle Vargas, Isabelle Carvalho, Fabiana Rodrigues de Góes, Thiago Pardo, and Fabrício Benevenuto. 2022. HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7174–7183, Marseille, France. European Language Resources Association.

Francielle Alves Vargas, Fabiana Rodrigues de Góes, Isabelle Carvalho, Fabrício Benevenuto, and Thiago A. S. Pardo. 2021. Contextual-lexicon approach for abusive language detection. In *Proceedings of*

the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), Held Online, 1-3September, 2021, pages 1438–1447. INCOMA Ltd.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Comput. Sci.*, 7:e598.