Exploring Portuguese Hate Speech Detection with Transformers

Gabriel Assis¹, Annie Amorim¹, Jonnathan Carvalho², Daniel de Oliveira¹, Daniela Vianna³ and Aline Paes¹

¹ Institute of Computing, Universidade Federal Fluminense, Niterói, RJ, Brazil
² Department of Informatics, Instituto Federal Fluminense, Itaperuna, RJ, Brazil
³ JusBrasil

{assisgabriel,annieamorim}@id.uff.br, joncarv@iff.edu.br,
{danielcmo,alinepaes}@ic.uff.br, dvianna@gmail.com

Abstract

Social Media platforms, vital for debate and communication, also grapple with misinformation and hateful comments. This work examines the detection of hate speech in Portuguese, contemplating its unique linguistic and cultural nuances. Leveraging Transformer-based models and different training and activation strategies, eight models with variations in architecture, size, and pre-training corpora are evaluated. Our findings show that, even though large generative models with enhanced prompts exhibited promising results, tuned small language models are still superior in addressing this task.

1 Introduction

Social Media platforms have become essential for debate and enabled unprecedented communication. However, they have also introduced significant challenges, such as spreading misinformation and the proliferation of hateful comments (Pelle et al., 2018; Aluru et al., 2020). In the interim, the Transformer (Vaswani et al., 2017) architecture has emerged, demonstrating state-of-the-art results in various scenarios, including classification problems (Fortuna and Nunes, 2018).

However, detecting hate speech remains an open issue, notably lacking resources in languages other than English, such as Portuguese (Jahan and Oussalah, 2023). The inherent characteristics of the language play a crucial role in this context, as the use of figures of speech and cultural nuances can significantly complicate this problem (Jang et al., 2023). On the other hand, an equally important consideration is the sensitivity of the domain, where both types of misclassification – falsely identifying content as problematic and failing to identify problematic content – are critical, as they could lead to censorship or a failure to protect vulnerable groups.

In this context, we approach the problem as: Given a social media post $P$ written in Portuguese, pre-process it returning $X$, and classify it as belonging to one of the three classes in $Y = \{\text{“hate speech”}, \text{“offensive” or \textquote{neutral}}\}$, where offensive comments encompass rude or insulting communication, and hate speech involves expressions of hate towards an individual or a group, rooted on characteristics like ethnicity and gender (Pelle et al., 2018; Vargas et al., 2021).

Our work aims to investigate the performance of the prominent Transformer architecture to tackle this critical task, thereby contributing to safeguarding a resilient and pluralistic environment on social media. We explore eight models varying in architecture, size, and the corpora on which they were pre-trained. Specifically, we consider three groups of models: (i) models based on the BERT (Devlin et al., 2019) architecture; (ii) Portuguese-language models based on the LLaMA (Touvron et al., 2023a,b) architecture; and (iii) general-purpose Large Language Models (LLMs). The first group consists of four models, peculiarly three alternatives specialized for Portuguese — including a model pre-trained on a corpus of tweets — and one multilingual alternative also pre-trained on tweets. The second group includes two 7-billion parameter models pre-trained on structured texts. Lastly, the third group involves one model from the popular GPT (Brown et al., 2020) family and the recently released Gemini-pro (Google, 2023), both not mainly pre-trained in Portuguese. This way, we contribute to a diverse study of models to address the challenging domain of detecting Portuguese hate speech on social media platforms.

2 Related Work

Identifying hate speech on social media has become a significant topic in recent years. Yet, the number of studies focusing on the peculiarities of the Portuguese language remains limited compared to English (Jahan and Oussalah, 2023). Some approaches address models based on BERT and its
state-of-the-art capabilities for classification tasks. In this context, (da Silva and Rosa, 2023) evaluated several distinct models, finding superior results in BERT-based models, such as BERTimbau (Souza et al., 2020), a finding reinforced by (Santos et al., 2022). (Jahan and Oussalah, 2023) present results indicating that language-specific models achieve better outcomes than multilingual alternatives.

Furthermore, LLMs and their remarkable abilities, are also being investigated for this task. (Assis et al., 2024) compare the GPT-3.5 and the Brazilian chatbot Maritalk\(^1\) with pt-BR BERT-based options, concluding that the latter group achieve better results. (Oliveira et al., 2024) contrast the same pair of LLMs, with a prompt engineering approach, and underscore Maritalk’s potential despite GPT’s higher performance. Additionally, (Chiu et al., 2022) assessed ChatGPT for detecting hate speech content, and (Nguyen et al., 2023) evaluated tuned LLaMA-2 models for detecting sexual, predatory, and abusive texts.

None of the aforementioned works conducted a study that comparatively includes the same vast amount of Portuguese-language models tuned as a ternary classification problem. Also, decoder models as the foundation for classifiers and a more recent LLM in an in-context learning (Brown et al., 2020) approach have not been evaluated either.

3 Method

This section details the selected models, training methods for classifier models, and inference strategies for generative models.

3.1 Encoder-based Classifiers Training

We select encoder-based models as follows. First, we have BERT-based models pre-trained with Brazilian Portuguese corpora: BERTimbau (Souza et al., 2020) in its large version, and AIBERTina (Rodrigues et al., 2023) in its 100m version, both pre-trained with more well-formed language; also BERTweet.BR (Carneiro, 2023), that is pre-trained with a corpus of tweets. Bernice (DeLucia et al., 2022) is also pre-trained on a Twitter data corpus, but it is multilingual. The most common fine-tuning strategy was adopted, stacking a classifier layer onto the language model and adjusting the model weights according to the training examples.

3.2 Decoder-based Classifiers Training

Regarding the decoder-based classifier models, Sabiá-7b-1 (Pires et al., 2023), which is built on the LLaMA-1 architecture, and Gervásio-7b-PTBR (Santos et al., 2024), built on the LLaMA-2 architecture, were selected. Both models were pre-trained on well-structured Portuguese text corpora. We used a tuning approach similar to the one usually adopted in encoder-based classifiers: stacking a classifier layer onto the language model. This choice stems from the decoder output of LLMs holding semantic meaning from the input, serving as text representations for classification tasks with prominent results (Li et al., 2023).

3.3 Generative LLMs Activation

The popular GPT-3.5-turbo (Ouyang et al., 2022) and the Gemini-pro 1.0 (Google, 2023), recognized for their remarkable performance in recent benchmarks, were chosen as generative large language models. Due to the constraints in adjusting the weights of these large and closed models, our strategy leverages their in-context learning capabilities by activating them with prompts (Brown et al., 2020). The responses’ effectiveness may be expressively influenced by how the prompts are crafted (White et al., 2023). This way, a well-known method involves embedding examples directly within the prompts. Our approach encompasses fixing the prompt instruction and exploring the choice of demonstrations and their impact on the models’ performance. The instruction is as follows: CLASSIFIQUE O TEXTO DE REDE SOCIAL COMO “DISCURSO DE ODIO” OU “OFENSIVO” OU “NEUTRO”. N TEXTO: target N Classe:\(^2\).

We rely on four ways to assemble prompts using examples: (a.) zero-shot, with no examples; (b.) one-shot, which includes a single example from one class; (c.) one-class-shot, which incorporates one example per class; and (d.) few-shot, which uses more than one example per class, precisely two in this study. For selecting examples, we introduced three strategies: (e.) random choice, (f.) based on semantic similarity, and (g.) based on the number of tokens. Strategies (f.) and (g.) start by sorting the set of demonstration candidates into clusters per class. They then pick examples close and far from the test instances’ embedding

---

1https://www.maritaca.ai/

\(^2\)In English that would be: Classify the social network text as “hate speech”, “offensive”, or “neutral”. N Text: target \n N Class:
representation or mode size. This way, we aim to evaluate how such extremes affect inference.

4 Experiments and Results

This section details the implementation process and presents the results obtained.

4.1 Experimental Setup

Models Setup All the fine-tuned models utilized an early stopping criterion for epoch selection and a batch size of 16. The encoder-based models had a learning rate of $10^{-5}$. The 7B models were fine-tuned using LoRA (Hu et al., 2022) strategy, with $r = 16$, $lora_alpha = 32$, and a learning rate of $10^{-4}$. Finally, the prompt-activated models had $\text{temperature} = 0.1$ and the max_token = 20.

Datasets Three datasets with hate content were used for evaluation. HateBR (Vargas et al., 2022), which includes comments gathered from the Instagram accounts of Brazilian politicians; OLID-BR (Trajano et al., 2023), featuring tweets and YouTube comments in Portuguese; and ToLD-BR (Leite et al., 2020), which consists of a collection of Brazilian tweets. All datasets were divided into 60% for training, 20% for validation, and 20% for testing. Dataset preprocessing involves anonymizing users with the @USER token, replacing URLs with HTTPURL token, and converting emojis into text.

4.2 Results

To address space limitations, we only included the best results for each model in Table 1, based on the F1-score for hate speech, which we conjecture is the most critical class. Encoder-based models, especially the one pre-trained on social media and Portuguese (i.e., BERTweet.BR), were the top performers in most datasets, suggesting that pre-training corpus is a crucial aspect.

Despite having more parameters, decoder-based 7-billion models were less successful than encoder-based models. This hints at a possible gap in their training on hateful content. Furthermore, the demonstration selection strategy for large models activated by prompts demonstrates potential. The GPT and Gemini models achieved most of their best results when selections were based on size or semantic similarity. They even outperformed models specifically adapted for Portuguese and further fine-tuned, Sabiá and Gervásio.

Overall, our findings emphasize the superiority of encoder-based models for this task. While generative models have shown potential, especially those trained on vast and diverse datasets, the targeted nature of encoder language models pre-trained on specific domains (e.g., social media and Portuguese) and adjusted explicitly for the task appears to be a critical feature for identifying hate speech.

5 Conclusions

We examined eight models with varying features derived from the Transformer architecture for hate speech detection task. Our findings indicate that despite the advanced abilities of generative LLMs, small models still play a crucial role in preventing the perpetuation of social issues in NLP tools. Additionally, aspects related to pre-training (e.g., the ethical filters, the nature and the language of the training corpora) may be correlated with better outcomes, more than the size of the models in this case. Our results also illustrate AI limitations for this critical domain. Therefore, we emphasize that these models should serve as aids in moderation, but not as complete substitutes for it.

Table 1: Macro results of precision, recall, accuracy, f1-score, and also the hate speech class f1-score for each model in its best configuration. Gemini* results may slightly fluctuate due to the rate of responses blocked by Google API filters. This rate was 0.15%, 0.79% and 0.12% for each dataset, respectively. Best results in bold.
Limitations

This study faces a limitation regarding the division of its training, validation and testing sets, as it employs only a single split. This constraint primarily stems from the significant costs of utilizing the GPT API, Gemini API, and computational resources on the Google Cloud environment. Furthermore, this limitation also restricted the variation of hyperparameters for our models, such as adjusting the number of epochs or modifying the learning parameters. While these factors may affect the interpretation of the models’ behavior in broader scenarios, those decisions enabled the analysis and comparison of various approaches across models, each with unique characteristics.

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. Additionally, this content was developed with the support of the Google Cloud Research Credits program, under the award GCP19980904.

References


Rodolfo Costa Cezar da Silva and Thierson Couto Rosa. 2023. Combining Data Transformation and Classification Approaches for Hate Speech Detection: A Comparative Study. Available at SSRN.


Zongxi Li, Xianming Li, Yuzhang Liu, Haoran Xie, Jing Li, Fu Lee Wang, Qing Li, and Xiaoqin Zhong. 2023. Label Supervised LLaMA Finetuning. CoRR, abs/2310.01208.


