Exploring Portuguese Hate Speech Detection in Low-Resource Settings: Lightly Tuning Encoder Models or In-Context Learning of Large Models?

Gabriel Assis[1], **Annie Amorim**[1], **Jonnathan Carvalho**[2],
Daniel de Oliveira[1], **Daniela Vianna**[3] **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, Manaus, AM, Brazil
{*assisgabriel,annieamorim*}*@id.uff.br*, *joncarv@iff.edu.br*,
{*danielcmo,alinepaes*}*@ic.uff.br*, *dvianna@gmail.com*

Abstract

Automatically identifying hate speech is an emerging field driven by the growth of social media and the consequent amplification of communication. However, this domain faces challenges due to the nuances of the language and variations in expression. In some countries, such as Brazil, the focus of this paper, hate speech can be typified as a crime by law. Nonetheless, enforcing the law is challenging, given the complexity of distinguishing hateful comments among the volume of interactions on social media. This work evaluates the abilities of language models to distinguish among neutral, offensive, and hate speech social media posts. Two classes of models are explored: three PT-BR BERT-based classifiers tailored explicitly for the task and two generative chatbots in an in-context learning approach. Given the impracticability of adjusting chatbots weights, we propose to enhance prompts by adding context based on topic modeling and selecting demonstration examples based on either their semantic or size proximity to the tested instances. The experimental results show that tuned small language models, even in a low-cost regime, are still superior to chatbots. Nevertheless, chatbots with enhanced prompts also exhibited promising results without further training.

1 Introduction

Social media are a powerful channel for disseminating information at an unprecedented speed, significantly enhancing the scope and capacity for communication and expressing opinions (Pelle et al., 2018). These platforms have evolved into virtual arenas for public debate, where individuals and groups can share their points of view on a wide range of topics (Moura, 2016; Paiva et al., 2019). However, these same platforms also magnify social issues such as the spread of misinformation, the proliferation of insults, and hate speeches (Aluru et al., 2020). In this context, offensive comments are defined as those containing any offensive communication, ranging from inappropriate language to direct insults (Pelle et al., 2018). Conversely, hate speech is characterized as any public expression of hate or violence encouragement towards an individual or a group based on characteristics such as ethnicity, race, nationality, sexual orientation, and gender (Vargas et al., 2021). Such expressions, when endorsed, potentially result in threats to individual integrity, thus emerging as a primary concern for digital communities, social media platforms, governmental entities, and society as a whole (Saraiva et al., 2021).

Moments of significant impact in public debate can make this task exceptionally challenging. For instance, in the federal-level elections in the United States in 2016, there was an increase in hate crimes (Edwards and Rushin, 2018). A similar effect was noticed in the 2018 Brazilian federal elections, when there was a massive increase in reports of xenophobia, homophobia, racism, and religious intolerance in social media (Vargas et al., 2021). Our work focuses on the Brazilian context. In Brazil, discrimination based on race, color, ethnicity, religion, or national origin is legally recognized as a crime[1]. Nevertheless, certain individuals misuse social media to disseminate such content, erroneously invoking the freedom of expression prerogative. While freedom of expression is a constitutional right, it must not promote hatred or intolerance. Nonetheless, applying the law remains a challenge, primarily due to the volume of posts and the complexity of identifying and classifying abusive comments (Vargas et al., 2021). Although digital platforms have their own prevention systems, they present several limitations. As an illustration, keyword filters can handle swear words, but not nuances in expressing hate (Yin and Zubiaga, 2021). Additionally, many users employ inventive tactics when writing offensive comments (Pelle

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

et al., 2018). This way, it is imperative to build accurate automated methods to filter and detect offensive and hate speech content. Thus, this paper tackles the following task: **Given a social media post P written in Portuguese, pre-process it returning X, and classify it as belonging to one of the classes in Y = {"hate speech", "offensive' or "neutral"}**.

Classifying social media posts is an active research field in Natural Language Processing (Fortuna and Nunes, 2018; Paiva et al., 2019; Jahan and Oussalah, 2023). In this vein, while the world is mesmerized by the generative chatbots remarkable abilities, like ChatGPT[2], tackling specifically challenging tasks such as identifying hate speech still remains. Nonetheless, adjusting the weights of these models is highly impractical due to their huge number of parameters, closed source code, and the implication of costs. The most viable alternative is to rely on in-context learning, wherein demonstrations are directly applied to prompts to incorporate context (Chiu et al., 2022).

In this paper, we evaluated various methods of demonstration selection: one-shot – which uses a single example regardless of the class – one-class-shot – with one example from each class – and the few-shot – which utilizes more than one example for each class. To select demonstration examples, we propose to choose them based on their size and similarity proximity to the test instances. We compare those strategies to select examples at random and not select any demonstration examples (zero-shot). Moreover, we propose to enhance the prompt context by adding keywords selected with topic modeling techniques while maintaining a fixed instruction.

However, a question that arises is if, even with enhanced prompts, chatbots are prepared to handle the specific language nuances to identify hate speech. In this sense, we investigate how classifiers based on relatively smaller models and minimally adjusted compare to the latest chatbots. We select encoder-based models given their significant results in classification (Fortuna and Nunes, 2018). Nonetheless, although adjusting the weights of such models is more feasible, other factors must be considered, like the characteristics of *corpora* they were trained, for example, style and text length.

Notably, we have three research questions regarding classifying social media posts as neutral, offensive, or hate speech, in two datasets[3].

- What is the performance of training low-cost classifiers from "small" language models? We employ minimal fine-tuning for only two epochs and train a classical classifier with feature extraction. We rely on the encoder-based models BERTimbau (Souza et al., 2020) and AlBERTina PT-BR (Rodrigues et al., 2023), trained with Brazilian Portuguese *corpora*. Moreover, given the tricky social media style, we add to the selection BERTweet.BR (Carneiro, 2023), trained with Brazilian Portuguese tweets.
- Does giving more context and fine-grained selected demonstration examples improve the response of chatbots? We compare the performance of two general-purpose chatbots with enhanced prompts, the popular ChatGPT (Brown et al., 2020) and MariTalk (Pires et al., 2023)[4] that is specifically trained with the Portuguese language.
- How do lightly adjusted BERT-based encoder models compare to general-purpose chatbots with enhanced context and examples? We conduct quantitative and qualitative investigations to shed light into the strengths and shortcomings of those models.

Our key findings and contributions are:

- Fine-tuning a tweets-based pre-trained small model prevails in detecting hate speech.
- Adding context and single well-selected examples benefits ChatGPT. Thus, this paper contributes with novel strategies for prompt enhancement that can be investigated in other domains and tasks.
- ChatGPT prevails over MariTalk in the hate speech and neutral classes, but not on the offensive class. While for ChatGPT, one-shot settings are the best options, MariTalk achieves two of its best results with zero-shot, pointing out less need for context.

2 Related Work

Although identifying hate speech in social media has become an imperative topic in recent years, the number of studies considering the peculiarities of the Portuguese language is still limited compared to English (Jahan and Oussalah, 2023; Trajano et al.,

[2] https://chat.openai.com/

[3] The code from our investigation is publicly available at https://github.com/MeLLL-UFF/hate_speech_in_context_pt

[4] https://chat.maritaca.ai/

2023). Nevertheless, some studies have applied and investigated traditional machine learning classifiers (da Silva et al., 2019; da Silva and Rosa, 2023; Paiva et al., 2019; Pelle et al., 2018; Plath et al., 2022; Souza et al., 2022; Vargas et al., 2021, 2022), Transformers (da Silva and Rosa, 2023; Leite et al., 2020; Oliveira et al., 2023; Plath et al., 2022; Santos et al., 2022; Vargas et al., 2021), and large language models (LLMs) (Chiu et al., 2022; Das et al., 2023; Oliveira et al., 2023) to address this issue.

In specific contexts, such as racism, misogyny, and homophobia, Naïve Bayes (NB), Support Vector Machines (SVM), and Random Forest (RF) classifiers trained with n-grams and bag-of-words have demonstrated good predictive performances (da Silva et al., 2019; Plath et al., 2022; Souza et al., 2022). In addition, some studies have used static embeddings (da Silva and Rosa, 2023; Pelle et al., 2018). However, such representations often limit the feature space regarding context-sensitive words.

BERT-based models have emerged as a prominent state-of-the-art in classifying hate speech, with some language-specific models outperforming multilingual alternatives in non-English contexts (Jahan and Oussalah, 2023). In this regard, da Silva and Rosa have evaluated 11 distinct classification methods, including BERTimbau (Souza et al., 2020), which has achieved the best results for the Portuguese language. Similarly, other studies highlighted the superior performance of BERTimbau (da Silva and Rosa, 2023; Santos et al., 2022) and multilingual BERT (Leite et al., 2020) over bag-of-words and static embeddings, considering hate speech as a binary classification problem.

Recently, general-purpose generative LLMs, such as GPT (Brown et al., 2020), have been analyzed in the task of hate speech and offensive text detection (Chiu et al., 2022; Das et al., 2023; Oliveira et al., 2023). Chiu et al. have investigated ChatGPT (Brown et al., 2020) for detecting sexist and racist language, employing zero-, one-, and few-shot learning techniques. In contrast, Oliveira et al. have exclusively employed the zero-shot technique to evaluate GPT's performance in hate speech detection in Portuguese tweets. The comparison with fine-tuned BERTimbau highlights the promising feasibility of GPT to classify hateful content. In (Das et al., 2023), ChatGPT has demonstrated good performance in hate speech detection for the Portuguese language, but it is limited in distinguishing counterspeech and non-hateful abusive speech targeting individuals and non-protected groups.

None of the aforementioned studies delve into pre-trained models for the social media environment or more recent encoder-based LMs for Portuguese with light tuning. Furthermore, regarding in-context learning of chatbots, no previous work has explored ways of enhancing the context with topic modeling or investigated demonstration examples of selection strategies. These features could prove to be insightful in the prompt construction phase. Moreover, the works relying on chatbots have not explored a chatbot trained for Portuguese.

3 Method

This section details the BERT-based and chatbot models adopted in this work, the training regimes applied to the BERT-based models, and the inference strategies proposed for the chatbots.

3.1 Models

We select BERT-based and large language chatbots models as follows. The BERT-based models are trained with Brazilian Portuguese *corpora*: (i.) BERTimbau (Souza et al., 2020) and (ii.) AlBERTina (Rodrigues et al., 2023) are trained with more well-formed language, and (iii.) BERTweet.BR (Carneiro, 2023) is trained with a *corpus* of tweets. The chatbots group includes (iv.) the popular ChatGPT, built upon GPT 3.5 (Brown et al., 2020) and (v.) MariTalk, built upon Sabiá LLM, tuned from GPT-based models and a Portuguese *corpus* (Pires et al., 2023). Given each model's size, nature, and open availability, we follow different evaluation regimes for those groups. However, the predictive performance is always measured from the same test set. The details come next.

3.2 Training regimes for BERT-based models

We trained the BERT-based models with two strategies: feature extraction and fine-tuning. Despite the usual higher predictive power of fine-tuning, we decided to also experiment with feature extraction because several previous works have followed this strategy to the hate speech detection task (Fortuna et al., 2019; Plath et al., 2022).

The feature extraction strategy selects the [CLS] token to serve as the input features to train SVM classifiers. In this case, only the classifier's parameters are adjusted to the training set, as the

pre-trained language model weights are frozen. We extract the feature vectors $\mathbf{X} \in \mathrm{R}^{n \times d}$ from the language models, where $\mathbf{X}$ is the examples matrix, n is the number of examples and d is the number of dimensions of token [CLS]. The other strategy is to stack a classifier layer to the language model and adjust the weights of the pre-trained language model according to the training examples, the most common fine-tuning setting.

3.3 Inference strategies for Chatbots

The answers from ChatGPT and MariTalk are gathered from their public APIs. Those agents receive as input a prompt composed of an instruction, a context, and zero or more demonstration examples. The instruction includes the task one wants the agent to perform, the context is any additional information provided, and an example is a pair (X, Y_i) to serve as a reference to the task. We tackle three classes in this paper, so Y_i can be neutral, offensive, or hate speech. This paper proposes several ways of selecting demonstration examples. Additionally, we also experiment with different contexts. We keep the instruction fixed. We are aware those agents are sensitive to the instructions. However, we rely on a previous study that explored instructions for hate speech detection in Portuguese (Oliveira et al., 2023). Complementary, we want to investigate the role of the context and demonstrations in composing the prompts.

3.3.1 Prompt

Two main resources inspire the instruction in this work. The first one is PromptHub[5], an open-source repository of prompts categorized by task. Prompts related to similar tasks, like sentiment analysis, from this collection helped shape the formulation of our instruction. On the other hand, Pires et al. influenced the integration of demonstrations within the prompts. Thus, we define the following instruction: CLASSIFIQUE O TEXTO DE REDE SOCIAL COMO "DISCURSO DE ODIO" OU "OFENSIVO" OU "NEUTRO".\N TEXTO: *target* \N CLASSE:[6].

3.3.2 Demonstration examples

Concerning the number of demonstration examples, we formulated four ways to compose the prompts: **(a.) zero-shot**, where no example is included in the prompt, **(b.) one-shot**, where a single example is included in the prompt, no matter its class, **(c.) one-class-shot**, where the prompt includes one example per class, and **(d.) few-shot**, where the prompt has more than one example per each class. All demonstration examples come from the training set.

To choose the demonstrations from the training set, we propose three strategies. The same examples are selected for all the test instances to account for less variability and more efficiency. The most straightforward strategy is **(e.) to select examples at random**, respecting the number of demonstration examples. For example, strategy (e.), together with (c.), chooses one example randomly from each class, while with (b.), it selects a single example from the whole training set. The two additional strategies consider either **(f.) the semantic similarity** according to the embedding representations or **(g.) the size in number of tokens** to select demonstration examples. Our intuition is to provide additional yet relevant information to better guide the in-context learning ability.

Both strategies start with automatically building clusters $C = \{C_{1,1}, \ldots, C_{1,k_1}, C_{2,1} \ldots, C_{2,k_2}, C_{3,1}, \ldots, C_{3,k_3}\}$ from the training set, separately for each one of the three classes, to account for better discernibility. In addition, they assume that all test instances belong to the same cluster C_t. Next, they identify the cluster $C_i \in C$ closest to the average embeddings of instances in C_t and the cluster $C_j \in C$ furthest to C_t. The intuition is to observe how the information on those extreme cases may contribute to or harm the in-context learning ability. To identify the clusters, we rely on the average distance of the examples of each $C_i \in C$ relative to C_t.

To further evaluate the role of extreme information, the **semantic similarity-based strategy** (f.) selects either (f.1.) the examples $Ex = \{ex_w \in C_i\}$ closest to the average embeddings of C_t according to the cosine similarity, or conversely, it selects (f.2.) the examples $Ex = \{ex_z \in C_j\}$ furthest to the average embeddings of C_t. Naturally, it must respect the a-d settings. For example, the few-shot case selects N examples, while the one-shot selects only one.

The **size-based strategy** (g.) builds upon (f.) by further selecting semantically close or distant examples that have a size most similar to the mode of the instances in the test set. This strategy comes

[5] https://github.com/deepset-ai/prompthub

[6] In English that would be: *Classify the social network text as "hate speech", "offensive", or "neutral". \n Text: target \n Class:*

from the observation that semantically close information might convey a similar amount of tokens to deliver similar messages.

3.3.3 Context

We experimented with two strategies: using no further context or including keywords to give the model examples of words representative of the type of discourse. Selecting keywords resembles the annotation task when the guidebook usually instructs the annotator to classify a text as hate speech or not, depending on the terms it contains (Vargas et al., 2022). Our proposed method consists of four steps. First, it removes possessive pronouns, proper nouns, verbs, stopwords, special characters, numerals, and words shorter than two letters from the instances. Next, for each class, it generates topics from the training set relying on BERTopic (Grootendorst, 2022) integrated with BERTimbau. Then, it counts the frequency of words for each class and marks the ten most frequent ones. Finally, it selects the ten most relevant words from the generated topics, provided they did not appear in the topics or the frequent word set of other classes.

The keywords are included in the prompt between the instruction and the demonstrations in the format: CONSIDERANDO QUE OS ASSUNTOS DA CLASSE "CLASS A" ESTÃO ASSOCIADOS COM AS PALAVRAS E EMOJIS *top 10 relevant terms in class A*. \N DA CLASSE "CLASS B" ESTÃO ASSOCIADOS COM AS PALAVRAS E EMOJIS *top 10 relevant terms in class B*. \N DA CLASSE "CLASS C" ESTÃO ASSOCIADOS COM AS PALAVRAS E EMOJIS *top 10 relevant terms in class C*.[7]

4 Experimental Setup

This section describes the experimental methodology and datasets used in the evaluation.

4.1 Experimental Methodology

The pre-processing procedure is straightforward, consisting of the removal of duplicates, replacing user mentions with the token @USER, links with HTTPURL, and emojis with their textual representation using the Emoji library[8]. Selecting clusters C as part of strategies (f.) and (g.), discussed in Section 3.3.3, relies on classical KMeans (Jin and Han, 2010). The number of clusters is selected according to the elbow criteria, and they were $k = 4$ for all classes. Training classifiers of Section 3.2 rely on default hyperparameters that come with the frameworks. Following a low-resource premise, we fine-tuned the BERT-based models for only two epochs with a learning rate of $2e-5$ and a batch of size 16. In the experimental setup for chatbots, we set the answer maximum token limit to 20 and disabled sampling. The temperature parameter was set to 0.1 for ChatGPT. For MariTalk, a slightly higher temperature of 0.3 was chosen to avoid generating empty responses observed with more restrictive values. Few-shot learning relies on two examples per class. We implemented the encoder-based models using Hugging Face's transformer framework (Wolf et al., 2020) in a Google Colaboratory[9] environment with limited availability of one Tesla T4 and one Tesla A100 GPU, the last used in the AlBERTina model only. Scikit-learn (Pedregosa et al., 2011) was used to train SVMs.

4.2 Datasets

The models were evaluated on two datasets, HateBR (Vargas et al., 2022) and ToLD-Br (Leite et al., 2020). HateBR comprises 7,000 Instagram comments collected from the profiles of Brazilian politicians in the second half of 2019. It comprises the following classes: hate speech (categorized as misogyny, fatphobia, xenophobia, etc.), offensive (but non-hate speech) texts, and non-offensive texts. Those are the three classes evaluated in this paper. A notable point is that only about 700 comments were labeled as hate speech.

The ToLD-Br dataset consists of 21,000 tweets collected between July and August 2019, labeled under the classes non-toxic, LGBTQ+phobia, obscene, insult, racism, misogyny, and xenophobia. Similarly, it is notable that only about 300 tweets were exclusively classified into a hate speech category. In this paper, posts classified as obscene and insulting form the offensive class, while the non-toxic category is considered as neutral, and the remaining classes form the hate speech class.

While HateBR was labeled by annotators who were at least Ph.D. candidates and experts in linguistics, hate speech, and computing, ToLD-Br did not have this educational level restriction for annotators. The two datasets explored criteria such

[7] In English that would be: *Considering that the subjects of class "class A" are associated with the words and emojis top 10 relevant terms in class A. \n From class "class B" they are associated with the words and emojis top 10 relevant terms in class B. \n From class "class C" they are associated with the words and emojis top 10 relevant terms in class C.*

[8] https://pypi.org/project/emoji/

[9] https://colab.google/

as gender diversity, political orientation, and race diversity among the annotators.

Both datasets were divided into 80% for training and 20% for test, keeping the proportion of the classes. We downsampled the majority class in the training set to account for balancing. Moreover, we also downsampled the examples in the test set to save costs when testing chatbots-based models.

5 Results

This section presents the results of classifiers and chatbots focusing on the hate speech class. Then, it includes an overall comparison of the best results for each class and a qualitative discussion.

5.1 Results of BERT-based models

Table 1 exhibits the results of predictive precision, recall, and f-measure concerning the hate speech class, and accuracy, to answer the first question elicited in the introduction.

The results show that BERTimbau performs better in the feature extraction strategy, while BERTweet.BR has the overall best results after fine-tuning, except for precision in HateBR and recall in ToLD-Br, when AlBERTina got better results. We were expecting that BERTweet.BR would perform better, given it was trained on tweets vocabulary. However, its feature extraction results were surprising, particularly for ToLD-Br. We conjecture that it might have an overfitted vocabulary representation that, when not facing any adjustment, could not cope well with a separate classification procedure to distinguish among different classes. The other models, on the other hand, did not face tweets during the intermediate masked language task, and that might have ended up helping them to aid SVM in distinguishing the different classes better. Despite being a more recent and larger model, AlBERTina did not achieve the overall best results besides those two mentioned before. However, as it is larger than the others, we would probably have to tune it for more epochs in more expensive hardware.

5.2 Inference with LLMs-based ChatBots

Tables 2 and 3 report the inference results using chatbots considering the same test sets as the previous section, aiming at answering our second research question. They focus on the demonstration strategies (a-g) discussed in Section 3.3.2.

Although MariTalk was trained from Portuguese *corpora*, ChatGPT still performs better. Unfortunately, we do not have further architectural or training set details of ChatGPT to add insights about possible reasons for that. Still, we noticed an interesting behavior: neither chatbot shows the same pattern comparing zero-shot and few-shot strategies. For example, ChatGPT is never better with the zero-shot setting, while MariTalk has two of the best results (precision and accuracy) in ToLD-Br with no demonstration examples. This can be related to the in-context ability requiring less prompt information when the model was trained in the same language as the task.

Few-shot based on semantic similarity benefits ToLD-Br in both chatbots, while in HateBR the best few-shot results are either with random or size-based examples. Overall, the one-class strategies (a single example or a single example per class) with semantically distance selection achieved better F1 results, pointing out that giving a well-selected example as demonstration might be enough to conduct the model weights to the appropriate places.

Next, Table 4 exhibits the results when adding keywords context to the prompts. We add that context only to the best results from the demonstration examples strategies, to observe if we can further improve in-context ability when giving additional information to the models besides the demonstration examples.

The precision results achieved by MariTalk are indeed improved with further context, making it reach the best results in both datasets. However, the enhanced context worsens all the other results for this chatbot. ChatGPT, on the other hand, benefits more from enhanced context, improving precision and accuracy for HateBR, and precision, accuracy, and F1 for ToLD-Br. Given that ChatGPT is not a model solely trained for Portuguese, its in-context ability benefits more from words guiding what the model should consider when completing the prompts. However, we can also observe that the recall for all cases is worse. Given that the metrics are computed for the hate class, we can conclude that, in general, topic words might guide the models to classify fewer instances as hate speech. This could be helpful to avoid incorrect censorship.

5.3 Comparative Results

This section presents comparative analyses regarding the best F1 result of each model for both datasets in Table 5, for the hate, offensive, and neutral classes, respectively. The previous results

	HateBR						ToLD-Br					
	Feature Extraction			Fine-tuning			Feature Extraction			Fine-tuning		
	BERTimbau	BERTweet.BR	AIBERTina	BERTimbau	BERTweet.BR	AIBERTina	BERTimbau	BERTweet.BR	AIBERTina	BERTimbau	BERTweet.BR	AIBERTina
prec.	**0.704**	0.401	0.623	0.761	0.777	**0.838**	**0.550**	0.000	0.429	0.647	**0.717**	0.438
rec.	**0.719**	0.568	0.691	**0.777**	**0.777**	0.597	**0.569**	0.000	0.466	0.569	0.655	**0.724**
acc.	**0.723**	0.406	0.683	0.800	**0.823**	0.771	**0.579**	0.320	0.433	0.652	**0.669**	0.534
f1	**0.712**	0.470	0.655	0.769	**0.777**	0.697	**0.559**	0.000	0.446	0.606	**0.685**	0.545

Table 1: Predictive results of feature extraction and fine-tuning-based classifiers. Except for the accuracy, they are computed to the hate speech class. Values in bold are the best for the category, while the best overall are underlined.

	ChatGPT										MariTalk									
	zero-shot	one-shot			one-class-shot			few-shot			zero-shot	one-shot			one-class-shot			few-shot		
		rand.	sim.	size	rand.	sim.	size	rand.	sim.	size		rand.	sim.	size	rand.	sim.	size	rand.	sim.	size
prec.	0.543	**0.600 (+10%)**	0.588 (+8%)	0.588 (+8%)	0.510 (-6%)	0.546 (+1%)	**0.615 (+13%)**	0.627 (+15%)	0.667 (+23%)	**0.691 (+27%)**	0.344	0.484 (+41%)	0.500 (+45%)	**0.573 (+67%)**	0.554 (+61%)	0.530 (+54%)	**0.577 (+68%)**	**0.655 (+90%)**	0.500 (+45%)	0.639 (+86%)
rec.	0.770	0.604 (-22%)	0.770 (=)	**0.799 (+4%)**	**0.906 (+18%)**	0.856 (+11%)	0.712 (-8%)	**0.640 (-17%)**	0.432 (-44%)	0.547 (-29%)	0.079	0.532 (+573%)	**0.568 (+619%)**	0.424 (+437%)	**0.734 (+829%)**	0.446 (+465%)	0.432 (+447%)	0.396 (+401%)	0.022 (-72%)	**0.561 (+610%)**
acc.	0.642	0.652 (+2%)	0.663 (+3%)	**0.675 (+5%)**	0.637 (-1%)	**0.688 (+7%)**	0.668 (+4%)	**0.695 (+8%)**	0.659 (+3%)	0.678 (+6%)	0.527	0.570 (+8%)	0.527 (=)	**0.594 (+13%)**	**0.652 (+24%)**	0.616 (+17%)	0.632 (+20%)	0.644 (+22%)	0.575 (+9%)	**0.678 (+29%)**
f1	0.637	0.602 (-5%)	**0.667 (+5%)**	0.657 (+3%)	0.653 (+3%)	**0.667 (+5%)**	0.660 (+4%)	**0.633 (-1%)**	0.524 (-18%)	0.610 (-4%)	0.129	0.507 (+293%)	**0.532 (+312%)**	0.488 (+278%)	**0.632 (+390%)**	0.484 (+275%)	0.494 (+283%)	0.493 (+282%)	0.041 (-68%)	**0.598 (+364%)**

Table 2: Inference Results of Chatbots in HateBR dataset considering different demonstration examples selection strategies. Except for the accuracy, they are computed to the hate speech class. Values in bold are the best for the category, while the best overall are underlined. The percentage in parentheses indicates the value compared to the respective zero-shot reference.

	ChatGPT										MariTalk									
	zero-shot	one-shot			one-class-shot			few-shot			zero-shot	one-shot			one-class-shot			few-shot		
		rand.	sim.	size	rand.	sim.	size	rand.	sim.	size		rand.	sim.	size	rand.	sim.	size	rand.	sim.	size
prec.	0.500	0.439 (-12%)	0.474 (-5%)	**0.495 (-1%)**	0.583 (+17%)	**0.588 (+18%)**	0.509 (+2%)	**0.778 (+56%)**	0.696 (+39%)	0.647 (+29%)	**0.857**	**0.750 (-12%)**	0.429 (-50%)	0.714 (-17%)	**0.800 (-7%)**	0.733 (-14%)	0.583 (-32%)	0.667 (-22%)	0.727 (-15%)	**0.778 (-9%)**
rec.	0.379	0.500 (+32%)	**0.621 (+64%)**	0.569 (+50%)	0.241 (-36%)	0.345 (-9%)	**0.483 (+27%)**	0.121 (-68%)	**0.276 (-27%)**	0.190 (-50%)	0.103	0.052 (-50%)	**0.103 (=)**	0.086 (-17%)	0.069 (-33%)	**0.190 (+84%)**	0.121 (+17%)	0.069 (-33%)	**0.138 (+34%)**	0.121 (+17%)
acc.	0.517	0.500 (-3%)	0.511 (-1%)	**0.528 (+2%)**	0.551 (+7%)	0.534 (+3%)	**0.552 (+7%)**	0.545 (+5%)	**0.562 (+9%)**	0.534 (+3%)	**0.562**	**0.478 (-15%)**	0.399 (-29%)	0.433 (-23%)	0.539 (-4%)	**0.556 (-1%)**	0.522 (-7%)	0.511 (-9%)	**0.545 (-3%)**	0.522 (-7%)
f1	0.431	0.468 (+9%)	**0.537 (+25%)**	0.512 (+19%)	0.341 (-21%)	0.435 (+1%)	**0.496 (+15%)**	0.209 (-52%)	**0.395 (-8%)**	0.293 (-32%)	0.185	0.097 (-48%)	**0.167 (-10%)**	0.154 (-17%)	0.127 (-31%)	**0.301 (+63%)**	0.200 (+8%)	0.125 (-32%)	**0.232 (+25%)**	0.209 (+13%)

Table 3: Inference Results of Chatbots in ToLD-Br dataset considering different demonstration examples selection strategies. Except for the accuracy, they are computed to the hate speech class. Values in bold are the best for the category, while the best overall are underlined. The percentage in parentheses indicates the value compared to the respective zero-shot reference.

	HateBR				ToLD-Br			
	ChatGPT		MariTalk		ChatGPT		MariTalk	
	no context	with context	no context	with context	no context	with context	no context	with context
prec.	0.588 (-7%)	**0.634**	0.554 (-16%)	**0.658**	0.474 (-31%)	**0.688**	0.733 (-27%)	**1.000**
rec.	**0.770**	0.597 (-22%)	**0.734**	0.540 (-26%)	**0.621**	0.569 (-8%)	**0.190**	0.138 (-27%)
acc.	0.663 (-5%)	**0.695**	**0.652**	0.637 (-2%)	0.511 (-13%)	**0.590**	**0.556**	0.517 (-7%)
f1	**0.667**	0.615 (-8%)	**0.632**	0.593 (-6%)	0.537 (-14%)	**0.623**	**0.301**	0.242 (-20%)

Table 4: Inference results when adding further context collected from word topics. We add context to the best results for F1 observed in Tables 2 and 3, namely one-shot for ChatGPT and one-class-shot for MariTalk. The best result for each chatbot is in bold, while the best result for each dataset is underlined. The percentage in parentheses indicates how much lower a value is compared to the best result.

did not include the other classes for two reasons. First, given its relevance and challenges, we wanted to give more visibility to the hate class. Additionally, the volume of data for the hate class is smaller than the others, hindering it if an average for all of them were presented. Second, presenting all the previous results would yield a volume of results incompatible with the paper limit of pages.

The results confirm the superiority of fine-tuned BERTweet.BR for the hate speech class, and also for the other classes in HateBR. In its benefits, BERTweet.BR was trained with a vocabulary of tweets, and social media platforms tend to share similar language styles. On the other hand, one would expect that its best results were in ToLD-Br, a tweets-based dataset. However, for the neutral and offensive classes this was not true; that might be related to the way this dataset was labeled. We give more details in the next section. Conversely, fine-tuned BERTimbau and ChatGPT with

one-class-shot setting and demo examples chosen at random got the best results for the offensive and neutral classes in ToLD-Br, respectively.

Focusing on the three best results for each class, we can confirm that BERT-based fine-tuned models prevail on most results for the three classes. Although this is an expected result, given they were fine-tuned with the datasets and chatbots were not, remember that their training was in a low-resource regime with only two learning epochs. Nevertheless, chatbots sometimes also appear in the three first positions of non-neutral classes – ChatGPT with an additional context in the hate class of ToLD-Br and zero-shot MariTalk in the offensive class of ToLD-Br. The neutral class presents the most divergent results: One-shot ChatGPT with demonstration examples based on size achieves the second-best result for HateBR and the best result for ToLD-Br with one-class-shot with random examples. Zero-shot MariTalk has the second-best result for this dataset. Given the low temperature set in their APIs, it could be the case that they are only returning the most likely answer. However, that might be a concern depending on how those chatbots are used in real-world applications and broader scenarios, as they might tend to overlook hate speech and offensive statements.

HateBR			ToLD-Br		
Rank	Model	F1	Rank	Model	F1
Hate Class					
2	BERTimbau (fine-tuned)	0.769	3	BERTimbau (fine-tuned)	0.606
1	BERTweet.BR (fine-tuned)	**0.777**	1	BERTweet.BR (fine-tuned)	**0.685**
3	AlBERTina (fine-tuned)	0.697	4	AlBERTina (fine-tuned)	0.545
4	ChatGPT (one-shot sim. hate)	0.667	2	ChatGPT (one-shot sim. off + ctx)	0.623
5	MariTalk (one-class-shot rand.)	0.632	5	MariTalk (one-class-shot sim.)	0.301
Offensive Class					
2	BERTimbau (fine-tuned)	0.775	1	BERTimbau (fine-tuned)	**0.687**
1	BERTweet.BR (fine-tuned)	**0.826**	2	BERTweet.BR (fine-tuned)	0.643
3	AlBERTina (fine-tuned)	0.756	5	AlBERTina (fine-tuned)	0.505
5	ChatGPT (one-shot sim. hate + ctx)	0.617	4	ChatGPT (one-shot rand. neu.)	0.580
4	MariTalk (few-shot size sim.)	0.637	3	MariTalk (zero-shot)	0.604
Neutral Class					
3	BERTimbau (fine-tuned)	0.857	4	BERTimbau (fine-tuned)	0.655
1	BERTweet.BR (fine-tuned)	**0.867**	3	BERTweet.BR (fine-tuned)	0.677
4	AlBERTina (fine-tuned)	0.853	5	AlBERTina (fine-tuned)	0.544
2	ChatGPT (one-shot size sim. neu.)	0.865	1	ChatGPT (one-class-shot rand.)	**0.710**
5	MariTalk (few-shot sim.)	0.835	2	MariTalk (zero-shot)	0.705

Table 5: Best F1 results per each class for each model and both datasets, highlighting the training regime or inference strategy and the model rank.

5.4 Qualitative analysis

There are four instances in HateBR where all models are misclassified. Two of them include a vulgar word with two letters referring to a part of the human body that might denote homophobic comments. Given the size of the word and the fact that it is a very common syllable, it might be the case that the models could not capture it as a token or its meaning. Another instance includes a possible sexist and fat-shaming comment ("PEPPAAAAA"), the name of an animated character that was commonly attributed to a female politician in Brazil[10]. It is quite likely that the subword-based tokenizers could not identify it as a token. Lastly, the other misclassified instance includes the 😂 emoji and a comment related to corruption, which does not seem to be a hate speech comment.

An instigating scenario is when all the BERT-based models correctly classify the instance as hate speech, but the chatbots do not: four out of five instances mention the left wing. Chatbots training might include reinforcement to avoid political partisanship, but that could also make them not capture that context. On the other hand, the chatbots correctly classify eight instances that BERT-based models do not; three are against a former Brazilian president, who is of the far-right, while two are in favor of him. A worrying pattern is observed in comments that include the word "NEGO" in ToLD-Br instances to denote anyone. There are six instances with that word labeled as hate speech, but all the models, except a single case for ChatGPT, mark them as neutral or offensive. While that word might not be used explicitly to denote racism, previous studies point out that it should be avoided given its historical adoption in racist terms (Guimarães Nascimento and Ribeiro, 2018). Those cases highlight the need for more extensive analysis of language models regarding social and political bias.

Finally, both datasets have some labels that caught our attention. For example, "TÔ CHEIA DE FOME E NADA DESSE ÔNIBUS VIR, QUE MERDA !", in English "I'M HUNGRY AND THIS BUS DOES NOT ARRIVE, SHIT" is labeled as hate speech in ToLD-Br. BERTimbau and ChatGPT classify it as neutral, while the others classify it as offensive. While it is a complaint, no offense is made. We also disagree with two other instances that chatbots agree with the annotation: "ESSE BOLSOLIXO É

[10] https://bit.ly/joice-hasselmann-e-peppa-pig

UM CANALHA ...” translated as “THIS BOLSOGARBAGE IS A BASTARD ...” offend the former president but do not attend hate speech criteria definition. Those cases show how challenging this task is, even for humans.

6 Conclusions

This paper investigates BERT-based models adapted to hate speech detection in PT-BR and different prompt adaptations for chatbots. We proposed two ways of enhancing prompts: adding topic-modeling context words and selecting demonstration examples to add more semantics to the demonstration. Selecting rich demonstration examples and including context benefits some of the chatbots settings. However, despite the recent increasing popularity of chatbots and their in-context abilities that claim no further training, we showed that adapting BERT-based models for those challenging datasets, even in a light training regime, still achieves the best results in most cases.

In this way, we reinforce the recent literature that argues for more investigation into the language model's abilities to handle sensitive social patterns such as hate speech, particularly in Portuguese. Small models still have a role in avoiding perpetuating social issues in NLP tools. Future investigation could focus on interpreting the role of layers, training *corpora*, and different architectural details in BERTimbau, BERTweet.BR and AlBERTina. Also, future work could further explore settings for our prompt enhancement proposals and see if they are helpful in other classification problems.

Limitations

This work presents some limitations concerning the division of training and tests. Firstly, there is only one split of the training and testing sets. Likewise, the adopted test set does not directly reflect the proportion of the classes observed in the real-world data sample. Both constraints arise mainly from the significant costs when using ChatGPT and limited request rate available via MariTalk. Another issue is the computational cost tied to the refinement of some adopted models, which involves adjusting up to 900 million parameters. Nevertheless, we had preliminary results employing cross-validation to the BERT-based models and HateBR dataset, when most of the results were similar to the ones presented in the paper. However, given the aforementioned costs and the need to be fair in comparing all the models with the same test sets, we presented the results without cross-validation procedures. This way, this paper assumes a low-resource scenario motivated by the need to reduce costs. Because of that, we do not explore other hyperparameters, such as temperature of chatbots and more epochs for BERT-based models. While these aspects may impact the interpretation of the models' behavior in more general scenarios, those decisions made it possible to analyze and compare several approaches across various models, each with its specific particularity.

Ethics Statement

Misclassifying offensive and hate speech content carries significant ethical implications and thus requires careful consideration and vigilance. Datasets may harbor cultural and historical biases, failing to encompass the full range of linguistic diversity. In this respect, Brazil is a prime example of cultural diversity; merely examining different perspectives within the same country can reveal discrepancies in the perceived offensiveness of a term. Additionally, when considering inter-country perspectives, such differences can become even more pronounced even among those speaking the same language. For instance, “*rapariga*” in Portugal primarily means “young woman”, while in Brazil, the term might carry derogatory connotations towards a woman[11]. Another critical point involves the potential hate speech false positives – especially in contexts where language use is ambiguous or employs figures of speech like irony and sarcasm – which could lead to unwarranted censorship by algorithms. Equally significant, false negatives for such classifications could fail to protect vulnerable groups and in the non-enforcement of laws. Therefore, we emphasize that AI mechanisms should serve as aids in content moderation, but should not be direct replacements for it.

Acknowledgements

This research was financed by CNPq (National Council for Scientific and Technological Development), grants 311275/2020-6 and 315750/2021-9, FAPERJ - *Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro*, process SEI-260003/000614/2023, and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

[11] bit.ly/rapariga-Brasil-Portugal

References

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Fernando Pereira Carneiro. 2023. BERTweet.BR: A Pre-Trained Language Model for Tweets in Portuguese. Master's thesis, Universidade Federal Fluminense, Programa de Pós-Graduação em Computação, Niterói.

Ke-Li Chiu, Annie Collins, and Rohan Alexander. 2022. Detecting hate speech with GPT-3.

Rodolfo Costa Cezar da Silva, Deborah Silva Alves Fernandes, and Márcio Giovane Cunha Fernandes. 2019. Classificaçao de mensagens em língua portuguesa com traços de racismo no twitter. *Revista de Sistemas de Informaçao da FSMA*, 23:2–9.

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.

Mithun Das, Saurabh Kumar Pandey, and Animesh Mukherjee. 2023. Evaluating ChatGPT's performance for multilingual and emoji-based hate speech detection. *CoRR*, abs/2305.13276.

Griffin Sims Edwards and Stephen Rushin. 2018. The Effect of President Trump's Election on Hate Crimes. *SSRN Electronic Journal*.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4).

Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. 2019. A hierarchically-labeled Portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104, Florence, Italy. Association for Computational Linguistics.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *CoRR*, abs/2203.05794.

Raquel Costa Guimarães Nascimento and Erislane Rodrigues Ribeiro. 2018. Uma análise discursiva dos memes "nego isso, nego aquilo". *Revista do Sell*, 7(1).

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

Xin Jin and Jiawei Han. 2010. *K-Means Clustering*, pages 563–564. Springer US, Boston, MA.

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.

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

Amanda Oliveira, Thiago Cecote, Pedro Silva, Jadson Gertrudes, Vander Freitas, and Eduardo Luz. 2023. How good is ChatGPT for detecting Hate Speech in Portuguese? In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 94–103, Porto Alegre, RS, Brasil. SBC.

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.

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

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.

Ramon Pires, Hugo Queiroz Abonizio, Thales Sales Almeida, and Rodrigo Frassetto Nogueira. 2023. Sabiá: Portuguese large language models. In *Intelligent Systems - 12th Brazilian Conference, BRACIS 2023, Belo Horizonte, Brazil, September 25-29, 2023, Proceedings, Part III*, volume 14197 of *Lecture Notes in Computer Science*, pages 226–240. Springer.

Hannah O. Plath, Maria Estela O. Paiva, Danielle L. Pinto, and Paula D. P. Costa. 2022. Detecção de

discurso de Ódio contra mulheres em textos em português brasileiro: Construção da base mina-br e modelo de classificação. *Revista Eletrônica de Iniciação Científica em Computação*, 20(3).

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-*.

Raquel Bento Santos, Bernardo Cunha Matos, Paula Carvalho, Fernando Batista, and Ricardo Ribeiro. 2022. Semi-Supervised Annotation of Portuguese Hate Speech Across Social Media Domains. In *11th Symposium on Languages, Applications and Technologies (SLATE 2022)*, volume 104 of *Open Access Series in Informatics (OASIcs)*, pages 11:1–11:14, Dagstuhl, Germany. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Ghivvago Damas Saraiva, Rafael Anchiêta, Francisco Assis Ricarte Neto, and Raimundo 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)*, pages 1261–1267, Held Online. INCOMA Ltd.

Andrey Souza, Eduardo Nakamura, and Fabíola Nakamura. 2022. Detecção de Discurso de Ódio: Homofobia. In *Anais do XVI Brazilian e-Science Workshop*, pages 73–80, Porto Alegre, RS, Brasil. SBC.

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

Douglas Trajano, Rafael H Bordini, and Renata Vieira. 2023. OLID-BR: offensive language identification dataset for Brazilian Portuguese. *Language Resources and Evaluation*, pages 1–27.

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 Vargas, Fabiana Rodrigues de Góes, Isabelle Carvalho, Fabrício Benevenuto, and Thiago Pardo. 2021. Contextual-lexicon approach for abusive language detection. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1438–1447, Held Online. INCOMA Ltd.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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