How Good Is ChatGPT For Detecting Hate Speech In Portuguese?

Amanda S. Oliveira[1], **Thiago C. Cecote**[1], **Pedro H. L. Silva**[1], **Jadson C. Gertrudes**[1], **Vander L. S. Freitas**[1], **Eduardo J. S. Luz**[1]

[1]Computing Department – Federal University of Ouro Preto

{amanda.oliveira2,thiago.cecote}@aluno.ufop.edu.br

{silvap,jadson.castro,vander.freitas,eduluz}@ufop.edu.br

Abstract. *This study evaluates OpenAI's ChatGPT, a large language model, for its efficacy in detecting hate speech in Portuguese tweets, comparing it with purpose-trained models. Despite incurring considerable computational costs, ChatGPT as a zero-shot classifier demonstrated commendable performance, even superior to or on par with state-of-the-art methods, with an F1-score of 73.0% on the ToLD-BR. In a cross-dataset evaluation on the HLPHSP dataset, it secured a superior F1-score of 73%. The choice of prompt significantly impacts the outcome, with a wider scope prompt balancing precision and recall metrics. ChatGPT, due to its interpretability and resilience against data distribution shifts, could be a preferred choice for tasks prioritizing these factors.*

1. Introduction

The considerable increase in users and time spent on social media has led to a growing prevalence of hate attacks, as the false sense of anonymity encourages individuals to post derogatory comments online. Hate speech encompasses messages that convey intolerance or aversion towards specific groups, such as ethnic, religious, sexual, or gender minorities and immigrants. It may include derogatory comments, threats, and incitements to violence [Walther 2022].

Given the importance of detecting hate speech on social media in recent years, particularly as more toxic posts on networks like Twitter tend to receive higher engagement [Salehabadi et al. 2022], it is crucial to implement content moderation rules to prevent the sharing of toxic content.

Hate speech detection on social media can be viewed as a text classification problem involving natural language processing (NLP) techniques and machine learning to identify and filter offensive messages. Researchers have addressed this issue in various languages, including Arabic [Mubarak et al. 2017], German [Wiegand et al. 2018], and English [Davidson et al. 2017, Zampieri et al. 2019, Mandl et al. 2019], highlighting the need to consider each language's structural and social aspects [Radfar et al. 2020]. This work focuses on the Portuguese language.

Advancements in Portuguese hate speech detection have been marked, with several studies contributing significantly. [de Pelle and Moreira 2017] compiled a corpus of 1,250 comments, mostly offensive, from a news site and established a benchmark using naive Bayes and SVM classifiers, the latter outperforming with a 77-82 F1-score. A European-Portuguese dataset of 5,668 tweets was developed by [Fortuna et al. 2019], labeled by

annotators of varying expertise using binary and hierarchical schemes. The dataset's utility was demonstrated using pre-trained embeddings and LSTM.

Further advancements include the categorization of 7,000 Instagram comments as hate speech/offensive or non-offensive by [Vargas et al. 2022], using n-grams and bag-of-n-grams with tf-idf preprocessing and achieving an F1-score of 85% for hate speech detection and 78% for offensive speech. [Leite et al. 2020] introduced a large-scale Brazilian Portuguese dataset, ToLD-Br, containing 21,000 annotated tweets. BERT-based models were applied, achieving a macro-F1 score of 76% in binary mode. Despite these developments, there is room for improvement of large-scale monolingual data.

With the growing public interest in generative pre-trained transformer (GPT) models, such as OpenAI's ChatGPT [Brown et al. 2020] [1], it is natural for this type of model to be employed for various natural language tasks, including toxicity analysis in social media texts. However, Large Language Models (LLM) like ChatGPT are designed to be generalists and function as a chatbot. In this work, we investigate three research questions: (i) As a chatbot model, can ChatGPT effectively detect hate/toxic speech in social media texts? (ii) How well does ChatGPT perform compared to models specifically trained for hate/toxicity detection tasks? And (iii) what is the impact of the prompt? To address these questions, we propose a methodology for classifying tweets as hate speech or non-hate using the ChatGPT API in a zero-shot classification fashion. Additionally, we compare ChatGPT's performance against a published baseline and a naive approach.

Our findings suggest the promising feasibility of employing ChatGPT to classify toxic/hateful textual content within Portuguese tweets. Furthermore, ChatGPT provides a more interpretable solution, a characteristic of considerable desirability, albeit at a significantly elevated computational cost. Hence, it may not represent the most optimal solution, given that alternative methods, characterized by simplicity and less computational intensity, have demonstrated comparable performance. The highest efficiency attained using ChatGPT was a macro-average F1-score of 73%, whereas the method based on BERTimbau-large generated a mean macro-average F1-score of 75%. However, in a cross-dataset scenario, ChatGPT presented better figures (macro-average F1-score of 74% vs. macro-average F1-score of 69%). This comparison underscores the need to consider a balance between computational cost and classification performance in selecting the appropriate model.

2. Methodology

2.1. ChatGPT as hate speech classifier

ChatGPT, developed by OpenAI, is a LLM utilizing the multi-layered Transformers architecture as its information processing units. We believe that the ease of use and accessibility of ChatGPT make its adoption inevitable for natural language processing tasks by non-specialized individuals.

This study employed the OpenAI API with the GPT-3.5-turbo model and the ChatCompletion module to classify social media texts as toxic/hate speech. As the module operates as a chat interface, the inclusion of two types of messages, a system message and a user message, was necessary to create the appropriate prompt. This study explores two

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

variations of user messages (prompt). One with a more narrow focus, which we refer to as Prompt #1, and another with a broader content message, which we refer to as Prompt #2. Since the queries are in Portuguese, find following the english translation for the reader reference:

System: "Você está realizando análise de texto." (You are performing text analysis.)

User Prompt #1: "Me diga, com sim ou não, se você considera esse texto tóxico: " (Tell me, with a simple 'yes' or 'no', whether you regard this text as toxic.) + text

User Prompt #2: "Responda apenas 'sim' ou 'não' caso o texto abaixo contenha discurso de ódio, ofensa, agressividade ou outras formas de toxicidade:" (Respond with a simple 'yes' or 'no' if the following text contains hate speech, offense, aggression, or any other forms of toxicity:) + text

In this case, the "text" refers to each comment (tweet) from the test dataset, which is processed individually. To illustrate, some prompts are structured as follows:

User: "Me diga, com sim ou não, se você considera esse texto tóxico: @user ridículo isso" (Tell me, with a simple 'yes' or 'no', whether you regard this text as toxic: @user this is ridiculous)

*User: "Me diga, com sim ou não, se você considera esse texto tóxico: @user neguinho tá cansado de apanhar e baixar a cabeça" (Tell me, with a simple 'yes' or 'no', whether you regard this text as toxic: @user the n****'s tired of taking hits and bowing his head)*

2.2. Baseline methods

For the baseline models, we followed the methodology proposed in [Leite et al. 2020] for BERT-based models since BERT-based models approximate the state-of-the-art for other languages [Zampieri et al. 2019] and Portuguese as well.

The process of classification with BERT-based models encompasses several steps. Initially, the input text undergoes tokenization, breaking it into subwords or WordPieces. Following this, the tokenized input is transformed into high-dimensional continuous representations in a process known as embedding. BERT enhances these token embeddings, integrating token, segment, and positional embeddings to generate a more contextually enriched representation.

The core of the BERT-based model comprises multiple layers of transformer blocks. To facilitate classification, a linear layer is appended atop the BERT model. In this procedure, an activation function is employed - softmax was the chosen function for the work at hand. Lastly, the model undergoes training on labeled data employing an appropriate loss function. In this instance, the binary cross-entropy loss was utilized. In the case of all BERT-based methodologies, the larger variant of the model was employed.

To more faithfully reproduce the work of [Leite et al. 2020], we use the *simpletransformers* library [2] with arguments set to default values. Three model versions were investigated using different language models: BERTimbau[3] [Souza et al. 2020], Distil-

[2]https://simpletransformers.ai/

[3]https://huggingface.co/neuralmind/bert-large-portuguese-cased

Bert[4] [Sanh et al. 2019], and BertPierreguillou[5] [Guillou 2021]. In each model, the language base was established on BERT and subjected to training across diverse databases and tasks. The DistilBert is a multilingual case.

A naive approach was also implemented using a non-sequential model for comparison, called Linear Model. Following the preprocessing, the text data was tokenized and transformed into sequences using a straightforward Tokenizer. These sequences were then padded to achieve a consistent length of 280 tokens, corresponding to the maximum length of a tweet, thereby ensuring a uniform input shape for the model. The architecture comprised an Embedding layer with 16-dimensional embeddings, a Flatten layer, a Dense layer containing 32 neurons with a ReLU activation function, and a Dense output layer featuring a single neuron with a sigmoid activation function.

The study focused on binary classification, and the metrics employed included precision, recall, F1-score per class, and macro-F1. For all experiments, data preprocessing consisted of removing links and anonymizing user mentions. All models were trained using binary cross-entropy loss and the Adam optimizer, adhering to standard practices in the field. The source code can be accessed at `https://github.com/ufopcsilab/ToxicSpeech-ChatGPT-STIL`.

2.3. Evaluation Metrics

The models are evaluated with regard to several metrics, including class-specific F1-score, precision, and recall, along with their macro and weighted variants. The "macro" version of these metrics calculates the metric independently for each class and then takes the average, treating all classes equally. At the same time, the "weighted" version calculates metrics for each class independently, but when it averages them, it uses a weight that depends on the number of instances for each class. Confusion matrices are also employed, offering a more visual depiction of model performance across different classes.

3. Results and Discussion

This section presents the datasets used in the experimental design and the results of the experiments aimed at answering the research questions.

3.1. Datasets

The ToLD-Br is the primary dataset employed for evaluating the methodology. In contrast, the HLPHSD is a supplementary dataset aiming for cross-dataset evaluation. Here we only used HLPHSD as a test dataset.

ToLD-Br: The work proposed in [Leite et al. 2020] presents an extensive dataset focused on detecting toxic language in Brazilian Portuguese. Collected from Twitter, the dataset is larger than others found in the literature, covering various demographic groups and considering different types of toxic language: LGBTQ+ phobia, obscenity, insults, racism, misogyny, and xenophobia. Tweets outside of these toxic categories were deemed non-toxic (or non-hateful). A stringent annotation criterion was employed, in which three volunteers independently classified each tweet. A total of 21,000 annotated tweets were

[4]https://huggingface.co/Davlan/distilbert-base-multilingual-cased-ner-hrl

[5]https://huggingface.co/pierreguillou/bert-large-cased-squad-v1.1-portuguese

selected to compose the dataset, out of which 9,255 were classified as toxic and 11,745 as non-toxic. Data collection took place over two months (July and August 2019). The authors divided the dataset into 80% for training and the remaining for testing using a stratified strategy.

HLPHSD: Proposed in [Fortuna et al. 2019], the HLPHSD consists of 5,668 tweets from 1,156 users collected from January to March 2017. Each tweet was initially labeled in a binary manner (hate vs. no hate) by non-expert volunteers. Subsequently, a second round of labeling was conducted by specialists, with each tweet receiving multiple labels, resulting in a hierarchical taxonomy. In total, 81 hate speech categories were identified. Cohen's Kappa [Gamer et al. 2012] was used to verify the agreement between annotators. The authors collected data from Brazilian and European profiles, making the dataset diverse regarding this criterion. A total of 31.5% of the tweets are annotated as hate speech.

3.2. Can ChatGPT effectively detect hate/toxic speech in social media texts?

Large Language Models like ChatGP incorporate a certain degree of randomness in their output generation. Thus, the same query to ChatGPT can yield slightly different results. The "temperature" parameter primarily controls this randomness during the inference process. A higher temperature produces more diverse and creative outputs, while a lower temperature leads to more focused and deterministic outputs. To obtain deterministic output from the OpenAI API for each tweet in the ToLD-Br test set, we set the temperature parameter to zero. This implies that every query consistently generates the exact same result. In Table 1, we present the results for the two types of prompts investigated.

Table 1. Classification using ChatGPT-3.5-Turbo on the ToID-BR test set with GPT Temperature set to zero.

	Prompt #1			**Prompt #2**		
	Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
No-hate	0.84	0.55	0.66	0.80	0.69	0.74
Hate	0.60	0.87	0.71	0.66	0.78	0.72
Macro Avg.	0.72	0.71	0.69	0.73	0.74	0.73
Weight Avg.	0.73	0.69	0.68	0.74	0.73	0.73

3.3. How well does ChatGPT perform compared to models specifically trained for hate/toxicity detection tasks?

The experiment was conducted five times for each baseline approach, altering the seed value to generate variability in the stochastic components. Figure 1 illustrates the comparative results between the baseline method and the classification outcomes achieved by ChatGPT-3.5 Turbo. Each baseline method underwent training for ten epochs, utilizing a learning rate of 3×10^{-5}, batch size of 8, the Adam optimizer, and a binary cross-entropy loss function. Figure 2 displays the confusion matrix for the best model based on BERT and the classification by ChatGPT-3.5 Turbo with Prompt #2.

3.4. How well does the baseline model perform on a cross-dataset evaluation?

In an effort to enhance the complexity of hate speech classification for baseline models, a cross-dataset scenario has been evaluated in this study. The HLPHSD, similar to the ToLD-BR dataset, was collected from Twitter, albeit at a distinct temporal point, and focused

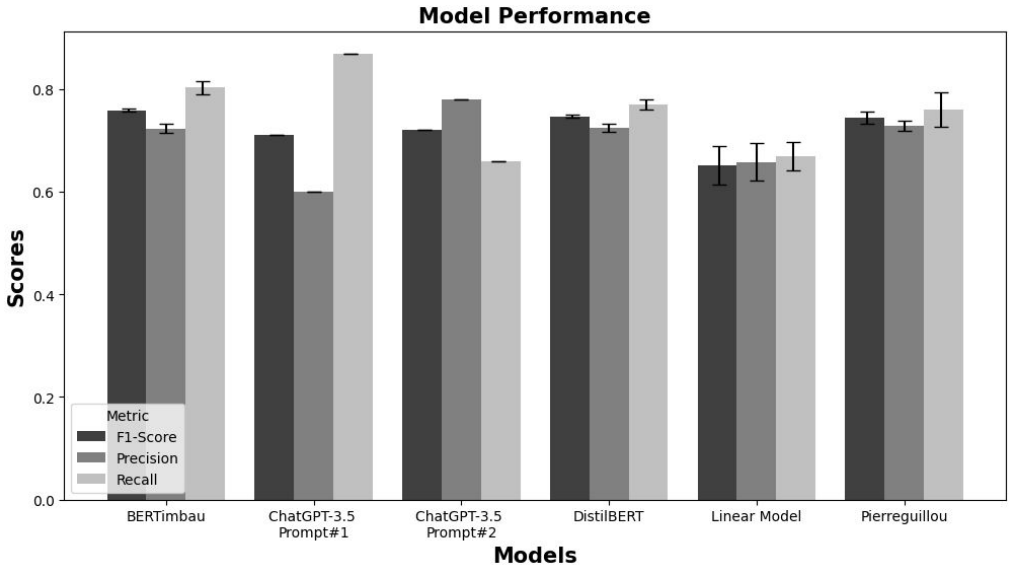
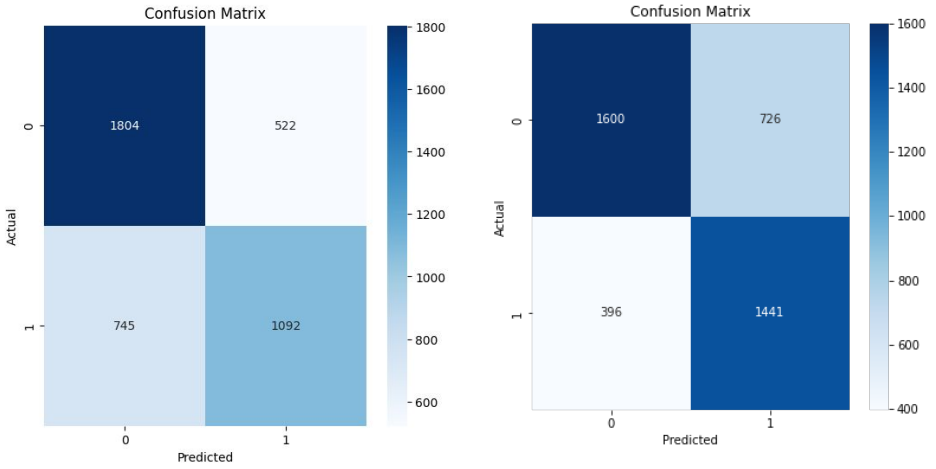


Figure 1. The results for precision, recall, and the F1-score pertaining to the hate speech class (1.0) within the ToLD-BR dataset are presented. Each experiment was conducted five times to ensure reliability.



(a) Confusion matrix for the model based on BERTimbau-large-cased.

(b) Confusion matrix for the ChatGPT-3.5 Turbo classification, with Prompt #2.

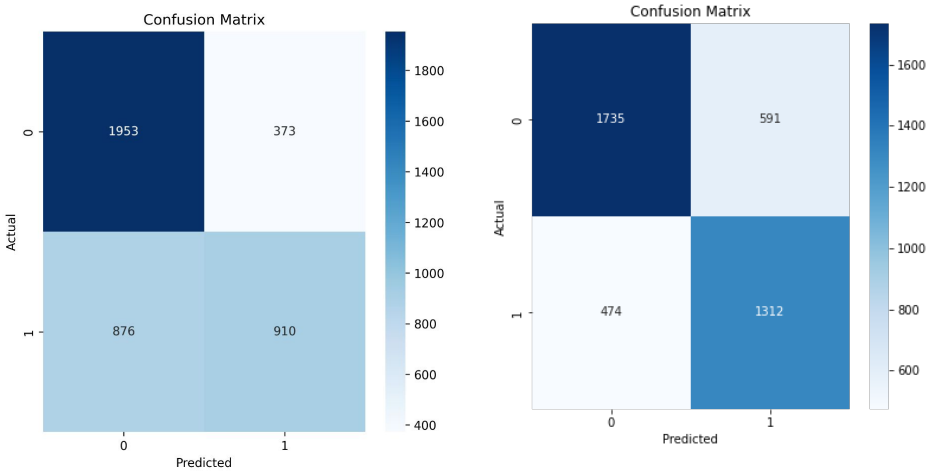
Figure 2. The confusion matrices from an experimental round on the ToLD-Br test dataset.

on Twitter profiles that utilized European Portuguese. The optimal model (employing BERTimbau) trained on the ToLD-BR dataset was selected for cross-evaluation. For this experiment, we balanced the HLPHSD dataset to align the distribution of instances per class with the ToLD-BR test set. The balancing consisted of randomly subsampling the majority class to equal the number of instances in the ToLD-BR test set, while retaining all samples from the minority class (hate speech). The results can be observed in Table 2. Also in Table 2, one can see the performance of ChatGPT-3.5 Turbo on the same dataset, the HLPHSD. Figure 3 displays the confusion matrix for the best model based on BERT

and the classification by ChatGPT-3.5 Turbo with Prompt #2.

Table 2. Cross-dataset results on the HLPHSD-balanced test set in terms of Precision (P), Recall (R) and F1-Score (F1).

	BERTimbau			ChatGPT Prompt#1			ChatGPT Prompt#2		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
No-hate	0.70	**0.82**	0.76	**0.81**	0.62	0.70	0.79	0.75	**0.77**
Hate	0.70	0.55	0.62	0.62	**0.81**	**0.71**	0.69	0.73	0.71
Macro Avg.	0.71	0.69	0.69	0.72	0.72	0.71	**0.74**	**0.74**	**0.74**
Weight Avg.	0.71	0.71	0.70	0.73	0.71	0.70	**0.74**	**0.74**	**0.74**



(a) Confusion matrix for the model based on BERTimbau-large-cased.

(b) Confusion matrix for the ChatGPT-3.5 Turbo classification, with Prompt #2.

Figure 3. The confusion matrices from an experimental round on the HLPHSD test dataset.

3.5. Discussion

Figure 1 clearly elucidates that ChatGPT-3.5 Turbo maintains a strong competitive stance, even without the employment of fine-tuning on the training data, hence exemplifying its capability for zero-shot classification. An important observation from the study is that the choice of the prompt dramatically influences the outcome. When the analysis utilizes Prompt #1, narrowly centered around the term 'toxic', it garners an exceptionally high recall albeit with a slight sacrifice in precision. In contrast, Prompt 2, characterized by its wider scope, yields a more harmonious balance between precision and recall metrics, albeit with a slight inclination towards precision. Notably, with the HLPHSD dataset - classified under a hate-speech-oriented taxonomy - Prompt #1 failed to provide any discernible advantages.

Furthermore, ChatGPT's distinct prowess is strikingly accentuated in the context of cross-evaluation, as evidenced by Table 2, as well as in Figure 3. Notably, the BERTimbau-cased-based model, originally trained with the ToLD-BR dataset, exhibited a decrease in performance during cross-evaluation. These findings decisively underline the exceptional abilities of ChatGPT-3.5 Turbo in its role as a Zero-shot Classifier for this task.

An intriguing observation emerged during experiments with ChatGPT, where the model not only classified tweets as “yes” or “no” for toxicity, but also autonomously provided justifications for its decisions, ranging from simple affirmations to sophisticated explanations. This indicates ChatGPT’s understanding of offensive content, which likely contributed to its effective ”toxic” categorization. However, its interpretations can skew when offensive language is used colloquially or rhetorically, leading to potential errors. Despite this, ChatGPT’s explanation capability could be invaluable in an industry setting, where model interpretability often holds significant importance.

Regarding computational cost, it is important to note that the ChatGPT-3.5 Turbo model has more than 175 billion parameters. In stark contrast, BERT-based models have an approximate parameter count of 100 million, and the simplest model, referred to here as the linear model, features roughly 1.74 million parameters. The employment of ChatGPT for this task will inevitably necessitate a significant energy outlay, thus entailing greater financial costs for inference.

4. Conclusion

The utilization of LLM-based models, such as ChatGPT, is observing an increasing surge of interest, with expectations pointing towards an expanded deployment in various NLP tasks, inclusive of text toxicity or hatefulness detection by both individuals and corporate entities. The insights gleaned from this study underline the competitiveness of ChatGPT-3.5 Turbo for this task, even when compared to models specifically fine-tuned for the same purpose. The results demonstrate a notable proficiency in toxic speech detection, boasting exceptional performance metrics and showcasing resilience towards shifts in data distribution. However, this prowess comes at a markedly higher computational cost when compared to other models assessed within this investigation. This highlights that, in practical terms, ChatGPT may not be the optimal selection for production use unless the requirement for interpretability is paramount. Smaller language models, such as Falcon [Penedo et al. 2023] and Llama [Touvron et al. 2023], can be an interesting future research path to balance computational cost, performance, and interpretability.

Acknowledgements

The authors would also like to thank *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* - Brazil (CAPES) - Finance Code 001, *Fundação de Amparo à Pesquisa do Estado de Minas Gerais* (FAPEMIG, grant APQ-01518-21), *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq, grant 308400/2022-4), and Universidade Federal de Ouro Preto (UFOP/PROPPI) for supporting the development of the present study.

References

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In Proceedings of the international AAAI conference on web and social media, volume 11, pages 512–515.

de Pelle, R. P. and Moreira, V. P. (2017). Offensive comments in the brazilian web: a dataset and baseline results. In Anais do VI Brazilian Workshop on Social Network Analysis and Mining. SBC.

Fortuna, P., da Silva, J. R., Wanner, L., Nunes, S., et al. (2019). A hierarchically-labeled portuguese hate speech dataset. In Proceedings of the third workshop on abusive language online, pages 94–104.

Gamer, M., Lemon, J., Gamer, M. M., Robinson, A., and Kendall's, W. (2012). Package 'irr'. Various coefficients of interrater reliability and agreement, 22:1–32.

Guillou, P. (2021). Portuguese bert large cased qa (question answering), finetuned on squad v1.1.

Leite, J. A., Silva, D., Bontcheva, K., and Scarton, C. (2020). Toxic language detection in social media for brazilian portuguese: New dataset and multilingual analysis. In 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.

Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., and Patel, A. (2019). Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In Proceedings of the 11th forum for information retrieval evaluation, pages 14–17.

Mubarak, H., Darwish, K., and Magdy, W. (2017). Abusive language detection on arabic social media. In Proceedings of the first workshop on abusive language online, pages 52–56.

Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E., and Launay, J. (2023). The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. arXiv preprint arXiv:2306.01116.

Radfar, B., Shivaram, K., and Culotta, A. (2020). Characterizing variation in toxic language by social context. In Proceedings of the international AAAI conference on web and social media, volume 14, pages 959–963.

Salehabadi, N., Groggel, A., Singhal, M., Roy, S. S., and Nilizadeh, S. (2022). User engagement and the toxicity of tweets. arXiv preprint arXiv:2211.03856.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9, pages 403–417. Springer.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Vargas, F., Carvalho, I., de Góes, F. R., Pardo, T., and Benevenuto, F. (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.

Walther, J. B. (2022). Social media and online hate. Current Opinion in Psychology, 45:101298.

Wiegand, M., Siegel, M., and Ruppenhofer, J. (2018). Overview of the germeval 2018 shared task on the identification of offensive language. In 14th Conference on Natural Language Processing - KONVENS 2018. Verlag der Österreichischen Akademie der Wissenschaften.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019). Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). arXiv preprint arXiv:1903.08983.