A Transformer-based Tabular Approach to Detect Toxic Comments

Ghivvago Damas[1][0000−0001−5466−6607], Rafael Torres Anchiêta[2][0000−0003−4209−9013], Raimundo Santos Moura[1][0000−0002−1558−3830], and Vinicius Ponte Machado[1][0000−0003−3391−8443]

[1] Federal University of Piauí, Brazil
{ghivvagodamas.ufpi,rsm,vinicius}@ufpi.edu.br
[2] Federal Institute of Piauí, Brazil
rta@ifpi.edu.br

Abstract. In recent years, there has been a significant increase in toxic and hateful speech on social media platforms, becoming deeply entrenched in online interactions. This issue has drawn the attention of researchers from various academic fields, leading them to extend their focus to include disciplines such as Natural Language Processing, Machine Learning, and Linguistics, in addition to traditional areas like Law, Sociology, Psychology, and Politics. This paper introduces an approach for detecting toxic and hateful speech on social media using Tabular Deep Learning. The goal is to apply and evaluate the performance of the FT-Transformer model in detecting hateful and toxic content in textual comments on social media in Brazilian Portuguese. An important aspect of this research involves using modern embedding models as language embedders and language models and evaluating their performance with the FT-Transformer, a transformer-based tabular model. The experimental scenario uses a binary version of the ToLD-Br dataset. Our approach achieved a 76% accuracy rate and a 75% macro F1-score using the OpenAI *text-embedding-3-large* model.

Keywords: Toxic and Hateful Speech · Deep Learning · FT-Transformer · Embedding models · Text Classification.

1 Introduction

With the rise of online social media platforms, increasing levels of toxicity and hateful behavior have become a growing concern. While social media has revolutionized communication itself in so many ways, it has also led to social issues beyond hate speech, including cyberbullying and misinformation. The anonymity of online platforms has created virtual public spaces where intolerance thrives, underscoring the need for strategies to address hateful speech and toxic content, as well as manage the negative impacts of social media use [39].

Detecting hate and toxicity in comments on social media platforms presents a complex challenge. Despite being identified as harmful, unlawful, or even criminal, these comments typically attract significant engagement and are frequently overlooked by the algorithms of social media networks [41], meaning that social media companies are fully aware of the nature of hate speech and its implications, however, the mechanisms applied to ensure the moderation policies are very limited or ineffective [43].

Researchers are developing advanced algorithms and strategies to effectively detect and filter hate speech and toxicity on social media to overcome these shortcomings in current moderation policies. Also, understanding the complexities of this issue is essential, including recognizing that not all offensive language necessarily constitutes hate speech or toxicity [18]. Factors such as context, the connection between speaker and audience, dynamics within social groups, the particular social media platform used, and timing play a role in accurately identifying harmful speech, including its subforms such as hateful speech [27]. Challenges include distinguishing between counter-hate speech, derogatory language used for emphasis, and subtler forms of hate and toxicity, such as sarcasm and metaphors [23]. The structure of language communication and its relationship with human perception are key factors in uncovering hate and toxicity online.

Due to the difficulties presented in this task, researchers are investigating different computational strategies to address it and have developed other approaches beyond full supervision. These strategies encompass a range of techniques from classical methods like Term Frequency — Inverse Document Frequency (TF-IDF) and Bag-of-Words (BoW) to vector representations, deep learning architectures, and Transformer-based models [27]. Modern methods have advanced to incorporate hybrid techniques and multimodal approaches [19], Language Models (LMs), Graph Machine Learning [37], and the use of Large Language Models (LLMs) [24] showing promising results.

Despite the various existing approaches, there is a lack of investigation into approaches for structured data, such as tabular data. This work looks at Tabular Deep Learning (TDL) as a viable and valuable tool for bringing diversity and unconventional alternatives into text classification tasks. Applying TDL models to hate speech detection can provide various benefits, including those found in the most commonly used models (SVM, XGBoost, LightGBM, and Neural Networks) and those that exceed its limitations when handling raw heterogeneous features, such as texts, sequences, images, audio, and embeddings. Integrating with other deep learning models or complex representation methods is also a limitation in most traditional methods, mainly in Gradient Boosting Decision Trees (GBDTs), in addition to low flexibility and inability to work with modern relational databases complexity and dimensionality [15].

This paper introduces an approach to detect hate speech and toxic comments in Brazilian Portuguese, employing the FT-Transformer (FTT) [13] model as a binary classifier. External text encoders, or Text Embedders (TE) [15], are applied to the FTT model. These TEs can be driven by modern embedding models and Pre-Trained Language Models (PTLMs) to transform raw or tokenized text into meaningful embedded representations known as embeddings. This transformer-based tabular approach was evaluated using the ToLD-Br [17] dataset, where various models were used to generate embeddings. Although the outcomes have not outperformed all hate speech detection methods, compared to other methods, this approach has shown competitive results as it requires fewer computational resources, allows faster training, and does not use any output optimization techniques.

The structure of this paper is as follows: Section 2 provides a concise overview of related work. Section 3 describes our developed approach in detail. In Section 4, we show the experimental scenarios and datasets, followed by an analysis of the results. Finally, Section 5 concludes the paper and outlines potential future research directions.

2 Related Work

The initial research on hate speech and offensive language paved the way for further exploration. Pioneering works by Chen et al. [8], Burnap and Williams [6], Waseem and Hovy [42], and Nobata et al. [22] defined and analyzed offensive and toxic speech, user behavior, and social media moderation.

For Brazilian Portuguese, significant contributions in hate, toxic and harmful speech detection, and text analysis were made by Almeida et al. [1], Pelle et al. [25], Bispo [4], Silva and Serapiao [33], Leite et al. [17], and Fortuna et al. [12]. Those studies applied machine learning techniques, textual feature representation methods, and discourse analysis to detect and analyze toxic and hateful comments on social media, advancing the field of automated hate speech detection.

The research by Almeida et al. [1] and Pelle et al. [25] expanded the work of offensive and harmful speech online. Almeida et al. [1] study proposed a hate speech identification strategy using Information Theory quantifiers, achieving an F1-score of 86%, 84%, and 96% for classifying hate, offensive, and regular speech classes. The study did not focus on hate speech in the Portuguese language. However, it still made valuable contributions to the scientific community and was useful for future Portuguese content research. Pelle et al. [25] introduced Hate2Vec, an ensemble-based classifier for detecting offensive comments on web platforms, performing well with datasets in English and Portuguese, compared to the traditional BoW classifier, attaining an F-score above 90%.

Further advancements in hate speech detection were made by Bispo [4] and Silva and Serapiao [33]. They developed classifiers using LSTM and CNN architectures, achieving significant accuracy with embeddings like Wang2Vec and GloVe. Bispo [4] specifically created a cross-lingual classifier from English to Portuguese using GBDTs, with F1-score accuracy ranging from 72% to 91%. In contrast, Silva and Serapiao [33] utilized a CNN architecture to identify hate speech, achieving F1-scores and Accuracy percentages between 82.64% and 96.74%. Different optimization techniques were employed for various datasets, such as Adam for OffComBR [10] and RMSprop for HLPHSD [11].

The work of Leite et al. [17] introduced the ToLD-br dataset for toxic comments in Brazilian Portuguese. It used fine-tuned BERTimbau and multilingual BERT for classification, achieving a macro F1-score of 76%. Fortuna et al.

[12] studied the generalization capabilities of classifiers for hate speech, toxicity, abusive, and offensive language, finding poor generalization with multilingual datasets and BERT, better generalization from English to Portuguese with a 67% macro F1 score, and the best generalization between English datasets with a 70% macro F1 score.

Saraiva et al. [32] introduced a novel semi-supervised node graph-based approach for detecting toxic comments with the ToLD-Br dataset. Their method utilizes an undirected and weighted Heterogeneous Graph Network (HGN) with 100-dimensional GloVe embeddings for the Portuguese language and achieves a 73% macro F-score using only 10% of the ToLD-Br. They used a Gradient Boost Classifier with Length-Generalization Consistency (LGC) as a transduction method.

Recent research has explored various machine learning and modern embedding models, such as Large Language Models (LLMs) embeddings and SBERT [28], for detecting toxic and hateful speech on social media. These models have proven effective in tasks like classification, clustering, and reranking in NLP. Studies have emphasized the value of modern embedding models in improving text-based tasks and their overall impact on NLP [20]. Additionally, research has shown the effectiveness of the Language-agnostic BERT Sentence Embedding model in cross-lingual and multilingual hate speech detection [31].

Other methods used different techniques with LLMs, including prompting and end-to-end classification. One study by Oliveira et al. [24] explored OpenAI ChatGPT (GPT-3.5-turbo) [5], the *ChatCompletion* module, and 2 prompts to assess GPT's performance compared to BERT-based models on various datasets. GPT achieved F1-scores of 73% as a zero-shot classifier on the ToLD-Br dataset and 74% when prompted in the cross-dataset experimental scenario with the balanced HLPHSD dataset. In another investigation by da Rocha Junqueira et al. [29], a cross-task comparison was performed between BERT-based models - BERTimbau and Albertina PT-BR. Despite fine-tuning efforts, Albertina PT-BR could not match the performance of BERTimbau. The base model of BERTimbau scored an F1-score of 88%, while its large model scored 89%. In contrast, the Albertina PT-BR base model only achieved a score of 74%.

The study by da Silva Oliveira et al. [34] examined the performance of LLMs like GPT-3.5-turbo and Maritaca AI Sabiá [26] in zero-shot and few-shot learning approaches, comparing them to the BERTimbau model. Sabiá demonstrated an enhanced and precise capacity to classify texts containing colloquial and slang expressions, and could identify aggressive and obscene words based on prompt-specific design. Analysis of the ToLD-Br revealed differences in performance between ChatGPT and MariTalk, with MariTalk showing improved precision due to its deeper understanding of Portuguese subtleties. Meanwhile, Assis et al. [2] conducted a study evaluating the ability of language models to distinguish neutral, offensive, and hateful speech in social media posts. The PT-BR BERT-based classifiers surpassed the chatbots on the HateBR [38] dataset, but on the neutral class, the chatbots outperformed the BERT-based classifiers on the ToLD-Br dataset. ChatGPT and MariTalk yielded F1 scores of 71% and 70%

respectively, which were lower than BERT-based classifiers, ranging from 77% to 86%.

Other initiatives dive into TDL models as alternatives for unconventional and novel binary text classification approaches and have gained attraction, with notable frameworks developed by Younus and Qureshi [44] and Chopra et al. [9]. The work of Chopra et al. [9] proposes an automated method for detecting hate speech in code-mixed Hindi-English text and Hindi text in Devanagari, where the framework architecture employs a TabNet classifier model trained on features extracted using a BERT-based model for Indian Languages (MuRIL) [16] from transliterated code-mixed data. This study demonstrated that the TabNet with MuRIL embeddings was effective for Devanagari text features, even though it was trained on transliterated data. The framework of Younus and Qureshi [44] highlighted the challenges of sexism detection and emphasized the importance of the number of training epochs in improving model performance, as ByT5 learns cleaner and finer representations.

Since TDL models are effective at handling complex dependencies and heterogeneity in tabular data, and modern embedding models have improved the quality of vector representations, introducing new methods for generating high-quality contextual embeddings the combination of these two can bring disruption to the hate speech detection scenario, and other NLP tasks.

Our research sets itself apart from previous methods by embarking on an initial investigation to address the lack of studies that integrate modern embedding models, embedding generation techniques, and TDL for detecting toxic and hateful comments on social media in Brazilian Portuguese. We emphasize the significance of testing different types of modern embeddings in the process of evaluating our classification strategy. TDL and modern embeddings are the core of our proposed method, which leverages the advantages of both in an attempt to achieve better outcomes.

3 Proposed Method

In this section, we propose a methodological approach for the binary classification of toxic and hateful comments on social media using Tabular Deep Learning (TDL). The developed approach applies the FT-Transformer (Feature Tokenizer + Transformer or FTT) model [13], slight adjustments on the base FTT model, integration to PyTorch Frame architecture for tabular data processing while incorporating language-specific embeddings for Brazilian Portuguese.

Figure 1 illustrates a five-step development process of this TDL approach, which includes: 1) Data preparation; 2) Embedding Generation; 3) Model Training; 4) Evaluation; and 5) Prediction. It is summarized as follows:

1) Data Preparation: The initial step involves gathering and normalizing the data to ensure it is suitable for training and has a consistent input format. This preparation removes unwanted data columns, shapes the data, and normalizes the text with Enelvo [3].

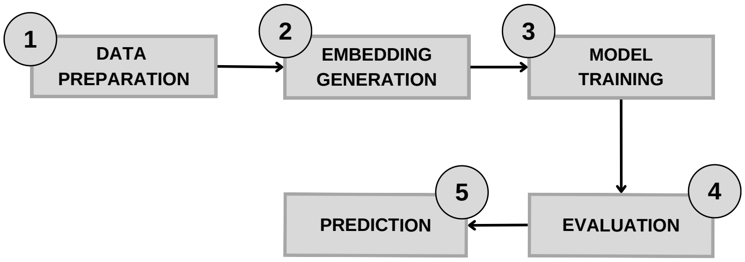

Fig. 1. A step-by-step process to detect toxic and hateful comments.

2) Embedding Generation: This step entails the textual data conversion into vector representations using a *Text Embedder* (TE), which is an LM-based embedding generation tool. The output representation matches the format and dimension from the input data and the embedding size of the model. To format the TE output into a tensor-friendly format suited for TDL pipelines, the *Materialization* step is required, where TE pre-encodes the text data before being shaped into a final tensor. The characteristics of the input text data and the language model used as an encoder in TE determine the final tensor shape.

3) Model Training: The processed data is then fed into the FT-Transformer model, which is subsequently trained. The FT-Transformer has two main components: the Feature Tokenizer (FT) and the Transformer. Since the original FT cannot ensure smooth operation and proper encoding of text data types in the dataset, this approach relies on its enhanced version reformulated by Hu et al. [15], *StypeEncoder*. After properly handling and processing each data type, the features are concatenated into a dense vector and passed to the transformer component as training input data.

4) Evaluation: After training, the model is evaluated using metrics such as F1-score and accuracy. This assessment helps in understanding the effectiveness of the approach in correctly detecting toxic comments while minimizing false positives and negatives.

5) Prediction: Finally, the trained model generates predictions on a separate test set. This step validates the model's capability to adapt and perform effectively on unseen data, demonstrating its robustness and reliability for other text classification tasks and real-world applications.

By following this structured approach, we aim to harness the capabilities of the FT-Transformer and advanced embedding models to improve the detection of toxic and hateful speech in social media contexts. Note that there is an intrinsic challenge in selecting suitable embedding models for tasks like detecting toxicity and hate speech in Portuguese comments since language models can introduce noise and impair the quality of the embeddings, leading to inferior performance.

So, to accommodate the language domain and its nuances, this approach also employs reliable Portuguese language models as *Generative Embedding Models* (or *Language Embedders*).

4 Experiments and Results

This section focuses on detecting toxic and hateful content in social media comments through binary text classification. The experimental scenarios described in this study outline the approaches used to conduct our research and assess the effectiveness of our proposed solution.

Our methodology is tested using the ToLD-Br [17] dataset, which consists of 21k social media comments. To ensure consistency and facilitate fair comparison with other methods, we divided the data into 80% for training, 10% for validation, and 10% for testing, in line with the original experiment. The selected model parameters for this experiment are listed in Table 1. Alongside using the AdamW optimizer, a scheduler was implemented to gradually reduce the learning rate from its starting value to zero during training.

Table 1. Model Parameters.

Parameter	Value	Parameter	Value
channels	256	out_channels	2
layers	12	learning_rate	0.0001
batch_size	512	loss function	'Cross Entropy'
epochs	100	optimizer	'AdamW'

The Text Embedder uses different models in each training round, these are: i) SBERT: E5-large [40] and SBERTimbau-large[3]; ii) BERT-based: BERTimbau [35], DeBERTa-V2-XL [14], and Albertina PT-BR [30]; and LLM embedding: VoyageAI (*voyage-large-2*) and OpenAI (*text-embedding-3-large*). Table 2 details these models, including their language support and the output embedding dimensions.

Table 2. List of Embedding Models.

Model Label	Language Support	Model Type	Output Dim
BERTimbau	Monolingual	BERT-based	1024
AlbertinaPTBR	Monolingual	BERT-based	1536
SBERTimbau	Monolingual	SBERT	1024
ME5Large	Multilingual	SBERT	1024
DeBERTaV2XL	Multilingual	BERT-based	1536
VoyageLarge2	Multilingual	LLMem	1536
OpenAI-TE3-large	Multilingual	LLMem	1536

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

The experiments performed in our study follow the least strict annotator agreement scenario described by Leite et al. [17]. In this agreement scenario, the dataset contains 11,745 non-toxic and 9,255 toxic comments, with a ratio in the class distribution of 1:1.2, which makes the ToLD-Br dataset can still be considered fairly balanced. After completing training rounds for each embedding model, a performance assessment can be done based on the obtained metrics. Table 3 displays *F1-score* for both toxic and non-toxic classes, as well as the overall accuracy of the chosen TE.

Table 3. Metrics Evaluation on different Text Embedders.

Text Embedder	**Toxic**	**Non-Toxic**	**Accuracy**
BERTimbau	0.7101	0.7531	0.7333
AlbertinaPTBR	0.6826	0.7306	0.6906
SBERTimbau	0.7158	0.7429	0.7300
ME5Large	0.7378	0.7258	0.7319
DeBERTaV2XL	0.6637	0.6974	0.6814
VoyageLarge2	0.7143	0.7327	0.7238
OpenAI-TE3-large	**0.7398**	**0.7740**	**0.7580**

Note: Toxic and Non-Toxic values represent the F1-score.

Based on these results, the OpenAI-TE3-large model has the highest accuracy and F1-score for the toxic class, making it the most robust choice. BERTimbau and SBERTIMBAU are strong performers among monolingual models, with BERTimbau being effective for the non-toxic class and SBERTIMBAU performing better for the toxic class. Multilingual embedding models like OpenAI-TE3-large, ME5Large, and VoyageLarge2 show excellent performance. Table 4 compares the results of different developed approaches used for detecting toxic and hateful speech using the ToLD-BR dataset and compares the overall F1-score and accuracy of each method.

Table 4. Toxic and Hateful Speech Detection: comparison of approaches using ToLD-BR.

Work	**Approach**	**F1-score**	**Accuracy**
[17]	M-BERT + *transfer learning*	0.76	-
[32]	GloVe + HGN + XGBoost	0.73	-
[24]	Prompt GPT 3.5	0.73	-
[29]	Albertina PT-Br Base	0.74	0.78
[29]	BERTimbau Base	0.88	0.88
[29]	BERTimbau Large	0.89	0.89
[34]	Prompt Sabiá + 10 few-shots	0.73	-
[34]	Prompt GPT 3.5-turbo + zero-shot	0.74	-
Ours	**OpenAI-TE3-large + FTT**	**0.75**	**0.76**
	BERTimbau + FTT	0.73	0.73
	ME5Large + FTT	0.73	0.73

The results show that the best-performing approaches for detecting hate and toxicity in social media comments involve extensive fine-tuning, transfer learning, or zero-shot learning, like BERTimbau Large and M-BERT as classifiers, and prompting GPT 3.5-turbo. However, methods that use graph representations, static embeddings, and modern embedding models can also achieve strong results even without additional tuning, providing a baseline for future improvements or hybrid implementations, and that was shown with our approach in the **OpenAI-TE3-large + FTT** setting. The confusion matrix depicted in Figure 2 serves as a subsequent assessment of the proposed approach, providing additional insights into the performance of the model over the classes (Toxic and Non-toxic).

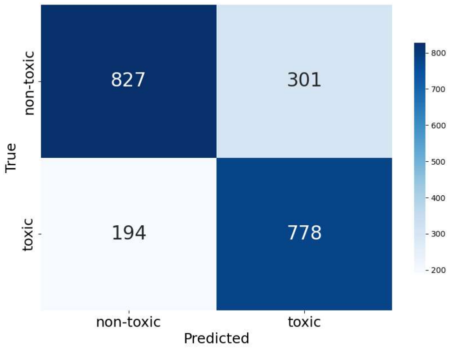

Fig. 2. Confusion Matrix for Toxic vs. Non-Toxic Classification.

The confusion matrix reveals that the model exhibits a robust ability to accurately identify non-toxic comments, as indicated by the high number of true negatives. Additionally, the model demonstrates a moderate capacity to detect toxic comments, with 778 true positives. However, the results also indicate notable misclassifications, with 301 false positives and 194 false negatives, suggesting room for improvement in reducing misclassification rates. In summary, the model has an overall accuracy of 76.43%, with a good balance between precision (72.08%) and recall (80.08%) for the toxic class.

Our Transformer-based tabular approach is available at `https://github.com/GhivvagoDamas/Tabular-Transformer-Toxic2024`.

5 Conclusion and Future Work

As previously demonstrated, we have proposed a novel approach for binary text classification, which is specifically tailored and trained for structured data in tabular format. This approach produced compelling results, successfully detecting hate speech and toxicity without complex fine-tuning or transfer learning methods. It yields an intriguing accuracy rate of 76% and an F1-score of 75% on the test set using the OpenAI *text-embedding-3-large* model as the Text Embedder.

Before the experiments, ToLD-Br underwent an additional review and evaluation, highlighting concerns about its validity and reliability. Issues were identified with the imbalance ratio in annotator agreement scenarios, possibly due to bias in the annotation process, term resignification, and the stylistic use of slurs.

Tabular Deep Learning models, like FT-Transformer, excel in multimodal learning and structured heterogeneous data processing with minimal adjustments. While the use of a GPU is important for this approach and implementations of other TDL models, no extensive computational resources are required. This study emphasizes the importance of TDL and Embedding Models in detecting toxic and hateful speech. Since modern embedding models have demonstrated enhanced processing efficiency and improved outcomes in different NLP applications [20], incorporating those in the proposed approach is not only adequate but also the most intelligent and assertive choice.

For future work, we intend to investigate more about modern embedding models like BGE M3 (FlagEmbeddings) [7], SBERT, and how to convert state-of-the-art LLMs like Sabiá [26], Aya [36], and other generative models into powerful and robust embedding models with refined embedding space and deep contextual understanding. Our intentions include adapting, and developing other TDL-based strategies and broadening NLP applicability, investigating bias detection and mitigation techniques similar to the study of Nascimento et al. [21], and model performance enhancement strategies such as incorporating additional contextual features, Contrastive Learning, and Retrieval-Augmented Generation (RAG).

Bibliography

[1] Almeida, T.G., Souza, B.À., Nakamura, F.G., Nakamura, E.F.: Detecting hate, offensive, and regular speech in short comments. In: Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web (2017)

[2] Assis, G., Amorim, A., Carvalho, J., de Oliveira, D., Vianna, D., Paes, A.: Exploring Portuguese hate speech detection in low-resource settings: Lightly tuning encoder models or in-context learning of large models? In: Proceedings of the 16th International Conference on Computational Processing of Portuguese, ACL, Spain (2024)

[3] Bertaglia, T.F.C., Nunes, M.d.G.V.: Exploring word embeddings for unsupervised textual user-generated content normalization. Proceedings of the 2nd Workshop on Noisy User-generated Text (2016)

[4] Bispo, T.D.: Arquitetura LSTM para classificação de discursos de ódio cross-lingual Inglês-PtBR. Master's thesis, Universidade Federal do Sergipe (2018)

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

[6] Burnap, P., Williams, M.L.: Us and them: identifying cyber hate on twitter across multiple protected characteristics. EPJ Data science **5** (2016)

[7] Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., Liu, Z.: Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation (2023)

[8] Chen, Y., Zhou, Y., Zhu, S., Xu, H.: Detecting offensive language in social media to protect adolescent online safety. In: 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing, IEEE, Amsterdam, Netherlands (2012)

[9] Chopra, A., Sharma, D.K., Jha, A., Ghosh, U.: A framework for online hate speech detection on code-mixed hindi-english text and hindi text in devanagari. ACM Transactions on Asian and Low-Resource Language Information Processing **22**(5) (2023)

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

[11] Fortuna, P., Rocha da Silva, J., Soler-Company, J., Wanner, L., Nunes, S.: A hierarchically-labeled Portuguese hate speech dataset. In: Proceedings of the Third Workshop on Abusive Language Online, ACL, Italy (2019)

[12] Fortuna, P., Soler-Company, J., Wanner, L.: How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? Information Processing & Management **58**(3) (2021)

[13] Gorishniy, Y., Rubachev, I., Khrulkov, V., Babenko, A.: Revisiting deep learning models for tabular data. Advances in Neural Information Processing Systems **34** (2021)

[14] He, P., Liu, X., Gao, J., Chen, W.: Deberta: Decoding-enhanced bert with disentangled attention. In: International Conference on Learning Representations (2021)

[15] Hu, W., Yuan, Y., Zhang, Z., Nitta, A., Cao, K., Kocijan, V., Leskovec, J., Fey, M.: Pytorch frame: A modular framework for multi-modal tabular learning. arXiv preprint arXiv:2404.00776 (2024)

[16] Khanuja, S., Bansal, D., Mehtani, S., Khosla, S., Dey, A., Gopalan, B., Margam, D.K., Aggarwal, P., Nagipogu, R.T., Dave, S., Gupta, S., Gali, S.C.B., Subramanian, V., Talukdar, P.P.: Muril: Multilingual representations for indian languages. CoRR **abs/2103.10730** (2021)

[17] Leite, J.A., Silva, D., Bontcheva, K., Scarton, C.: 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, ACL, China (2020)

[18] Malmasi, S., Zampieri, M.: Challenges in discriminating profanity from hate speech. Journal of Experimental & Theoretical Artificial Intelligence **30**(2) (2018)

[19] Mandal, A., Roy, G., Barman, A., Dutta, I., Naskar, S.K.: Attentive fusion: A transformer-based approach to multimodal hate speech detection. arXiv preprint arXiv:2401.10653 (2024)

[20] Muennighoff, N., Tazi, N., Magne, L., Reimers, N.: MTEB: Massive text embedding benchmark. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, ACL, Dubrovnik, Croatia (2023), https://doi.org/10.18653/v1/2023.eacl-main.148

[21] Nascimento, F.R., Cavalcanti, G.D., Da Costa-Abreu, M.: Unintended bias evaluation: An analysis of hate speech detection and gender bias mitigation on social media using ensemble learning. Expert Systems with Applications **201** (2022)

[22] Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., Chang, Y.: Abusive language detection in online user content. In: Proceedings of the 25th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, Montréal, Québec, Canada (2016), https://doi.org/10.1145/2872427.2883062

[23] Ocampo, N., Sviridova, E., Cabrio, E., Villata, S.: An in-depth analysis of implicit and subtle hate speech messages (2023), https://doi.org/10.18653/v1/2023.eacl-main.147

[24] Oliveira, A., Cecote, T., Silva, P., Gertrudes, J., Freitas, V., Luz, E.: How good is chatgpt for detecting hate speech in portuguese? In: Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana, SBC, Belo Horizonte/MG (2023)

[25] Pelle, R., Alcântara, C., Moreira, V.P.: A classifier ensemble for offensive text detection. In: Proceedings of the 24th Brazilian Symposium on Multimedia and the Web (2018)

[26] Pires, R., Abonizio, H., Almeida, T.S., Nogueira, R.: Sabiá: Portuguese large language models. In: Intelligent Systems, Springer Nature Switzerland (2023), ISBN 978-3-031-45392-2

[27] Rawat, A., Kumar, S., Samant, S.S.: Hate speech detection in social media: Techniques, recent trends, and future challenges. Wiley Interdisciplinary Reviews: Computational Statistics **16**(2) (2024)

[28] Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084 (2019)

[29] da Rocha Junqueira, J., Junior, C.L., Silva, F.L.V., Côrrea, U.B., de Freitas, L.A.: Albertina in action: An investigation of its abilities in aspect extraction, hate speech detection, irony detection, and question-answering. In: Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana, SBC (2023)

[30] Rodrigues, J., Gomes, L., Silva, J., Branco, A., Santos, R., Cardoso, H.L., Osório, T.: Advancing neural encoding of portuguese with transformer albertina pt-* (2023)

[31] Rodríguez, S.E., Allende-Cid, H., Allende, H.: Detecting hate speech in cross-lingual and multi-lingual settings using language agnostic representations. In: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 25th Iberoamerican Congress, CIARP 2021, Springer (2021)

[32] Saraiva, G.D., Anchiêta, R., Neto, F.A.R., Moura, R.: A semi-supervised approach to detect toxic comments. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), INCOMA Ltd., Held Online (2021)

[33] Silva, S., Serapiao, A.: Detecção de discurso de ódio em português usando cnn combinada a vetores de palavras. In: Proceedings of KDMILE 2018, Symposium on Knowledge Discovery, Mining and Learning, São Paulo, SP, Brazil (2018)

[34] da Silva Oliveira, A., de Carvalho Cecote, T., Alvarenga, J.P.R., de Souza Freitas, V.L., da Silva Luz, E.J.: Toxic speech detection in Portuguese: A comparative study of large language models. In: Proceedings of the 16th International Conference on Computational Processing of Portuguese, ACL, Spain (2024)

[35] Souza, F., Nogueira, R., Lotufo, R.: BERTimbau: pretrained BERT models for Brazilian Portuguese. In: 9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear) (2020)

[36] Üstün, A., Aryabumi, V., Yong, Z.X., Ko, W.Y., D'souza, D., Onilude, G., Bhandari, N., Singh, S., Ooi, H.L., Kayid, A., et al.: Aya model: An instruction finetuned open-access multilingual language model. arXiv preprint arXiv:2402.07827 (2024)

[37] Utku, A., Can, U., Aslan, S.: Detection of hateful twitter users with graph convolutional network model. Earth Science Informatics **16**(1) (2023)

[38] Vargas, F., Carvalho, I., Rodrigues de Góes, F., Pardo, T., Benevenuto, F.: 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, European Language Resources Association, France (2022)

[39] Walther, J.B.: Social media and online hate. Current Opinion in Psychology **45** (2022)

[40] Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., Wei, F.: Multilingual e5 text embeddings: A technical report. arXiv preprint arXiv:2402.05672 (2024)

[41] Wang, X., Koneru, S., Venkit, P.N., Frischmann, B., Rajtmajer, S.: The unappreciated role of intent in algorithmic moderation of social media content. arXiv preprint arXiv:2405.11030 (2024)

[42] Waseem, Z., Hovy, D.: Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In: Proceedings of the NAACL Student Research Workshop (2016), https://doi.org/10.18653/v1/n16-2013

[43] Yin, W., Zubiaga, A.: Towards generalisable hate speech detection: a review on obstacles and solutions. PeerJ Computer Science **7** (2021)

[44] Younus, A., Qureshi, M.A.: A framework for sexism detection on social media via byt5 and tabnet (2022)