Investigating Contextual Word Embeddings in Semi-Supervised Learning for Toxic Comment Detection

Francisco Assis Ricarte Neto
farn@ifpi.edu.br
Instituto Federal do Piauí
Universidade Federal do Piauí

Rafael Torres Anchiêta
rafael.torres@ifma.edu.br
Instituto Federal do Maranhão

Raimundo Santos Moura
rsm@ufpi.edu.br
Universidade Federal do Piauí

Pedro de A. dos Santos Neto
pasn@ufpi.edu.br
Universidade Federal do Piauí

Andre Macedo Santana
andremacedo@ufpi.edu.br
Universidade Federal do Piauí

ABSTRACT

The proliferation of toxic messages on the Web has led many social networks to limit or shut down user comments, as these messages are harmful to people and keep them away from online platforms. Most approaches deal with this unacceptable language form, focusing only on identifying if a comment is toxic or not, using a supervised learning strategy, leaving aside the various types of harmful messages, such as LGBT+phobia, racism, xenophobia, and others. In this paper, we investigate three contextual word embedding models: BERTimbau, BERTweet.BR, and LLaMA 3.1 within a semi-supervised approach to distinguish whether a comment is toxic. Also, we explore this strategy to detect six types of toxicity: LGBT+phobia, insult, racism, obscenity, xenophobia, and misogyny. This task is defined as a multi-label classification problem, as a comment may contain several types of abusive language. We evaluated our approach using the ToLD-BR corpus and achieved competitive results for the binary toxicity classification task. In the context of multi-label toxicity detection, our best result outperformed approaches based on supervised learning, using significantly fewer labeled data, and emphasized their efficiency and practicality.

KEYWORDS

Toxic Comments, Toxicity detection, Portuguese language, Multi-Label Classification

1 INTRODUCTION

Social networks are powerful tools for virtual human interactions, connecting people worldwide. These tools allow individuals to share information, express their opinions, and engage in debates. However, this scenario also attracts users who exceed the limits of freedom of expression, taking advantage of the space to disseminate all types of toxic language.

Toxic messages involve the use of inappropriate language that is considered unacceptable, including both explicit and implicit forms of profanity, insults, and threats directed at individuals or groups [41]. Toxicity can also manifest as negative behaviors, such as offensive comments or disrespectful remarks, hate speech, or any other characteristic that may deter a person from engaging in a conversation. Thus, fighting this language is of the utmost importance, as it is a crime in several countries. In this paper, we consider any comment with these characteristics to be toxic language.

Several strategies have been developed for toxicity detection, although most focus on the English language and rely on supervised machine learning techniques [28]. Recent efforts aimed at addressing toxicity in the Portuguese language have predominantly approached the task as a binary classification problem, extracting features from Bag-of-Words [26], word embeddings [35], and lexicons [39]. However, these methods typically depend on large annotated datasets to train effective classifier models. On the other hand, some studies have explored the use of Large Language Models (LLMs), which require less data for the training stage [2, 21, 22], though these approaches are often limited by their high computational and financial costs.

While these methods help to detect and combat unacceptable comments, they are not able to identify toxic groups such as sexism, racism, homophobia, xenophobia, misogyny, LGBTQ+phobia, and others. In an abusive language scenario, a comment may be on multiple toxic groups. For example, the sentence "*fu**ing could it feels like I am inside your heart*" may be considered both obscene and insulting. This is a multi-label problem, where a comment may be assigned to multiple labels. It is more challenging and less explored than the binary classification problem.

In this paper, we address the challenge of binary and multi-label classification of abusive language in Brazilian Portuguese, aiming to develop a solution that requires minimal labeled data. To this end, we adopt the semi-supervised approach proposed by Saraiva et al. [31], which models toxic comments as a heterogeneous graph structure. Heterogeneous graphs are robust data structures that can represent both simple and complex relationships in data-driven applications. Their main advantage lies in making explicit the connections between different types of objects [32]. In the model proposed by Saraiva et al. [31], toxic comments are represented as graphs composed of two distinct node types: sentence nodes and token nodes. These nodes are connected by undirected, weighted edges, where the weights are derived from the token embedding vectors $E(t)$. Figure 1 illustrates how a toxic comment is structurally represented within this heterogeneous graph framework.

Building upon this foundation, the original approach employs the Local and Global Consistency (LGC) algorithm [43] to propagate labels from a small set of labeled nodes throughout the graph, enabling the extraction of features from the graph topology and

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2025). Rio de Janeiro, Brazil. Porto Alegre: Brazilian Computer Society, 2025.

ISSN 2966-2753

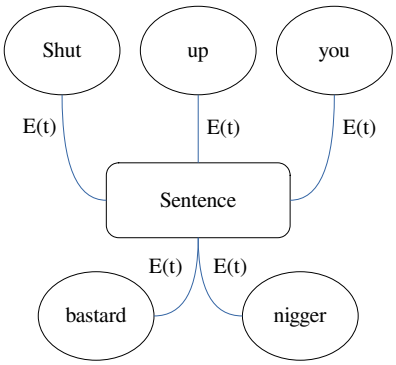

Figure 1: Example of a graph model for the sentence "*Shut up, you bastard nigger.*".

supporting the classification of all nodes. In this work, we extend the original proposal by incorporating an alternative regularization algorithm, Gaussian Fields and Harmonic Functions (GFHF) [44], to propagate labels and extract features from the graph structure. Unlike LGC, GFHF does not alter the labels of pre-annotated data during the regularization process, thereby preserving the integrity of the pre-labeled examples throughout propagation. This property is particularly important in low-data scenarios, as it ensures greater consistency and reliability during label dissemination.

In addition to modifying the label propagation strategy, we also improve the method used to compute edge weights. While Saraiva et al. [31] relied exclusively on static distributional word embeddings to assign weights to the edges connecting token nodes to sentence nodes, our approach employs three contextual word embedding models, BERTimbau [36], BERTweet.BR [4], and LLaMA 3.1 [10] to perform this task. Contextual word embeddings are more effective at capturing word meaning based on surrounding context. Unlike static embeddings, which assign the same vector to a word regardless of usage, contextual models generate distinct embedding vectors for each word occurrence depending on its context. As such, the resulting vectors can provide more accurate edge weight representations, as they are capable of encoding subtle linguistic nuances [34].

We evaluated our approach using the ToLD-BR corpus [16], which contains twenty-one thousand manually annotated comments across seven toxicity categories: *non-toxic*, *LGBTQ+phobia*, *obscene*, *insult*, *racism*, *misogyny*, and *xenophobia*. We also compared our method against supervised learning algorithms [16], LLM-based approaches [21, 22], and the semi-supervised method proposed by Saraiva et al. [31]. Using only 10% of labeled data for binary toxic comment classification, our method achieved competitive results compared to the supervised learning algorithms and outperformed both the LLM-based approaches and the semi-supervised method. Moreover, when using 30% of labeled data, our strategy surpassed supervised baselines based on bag-of-words and BERT [16] in the multi-label toxicity classification task. These findings suggest that contextual embeddings are more effective in capturing subtle linguistic nuances, which are crucial for accurately detecting toxic language within heterogeneous graph structures.

The remainder of this paper is structured as follows: Section 2 briefly describes related work. In Section 3, we present the corpus used. Section 4 details our approach to detecting toxic comments. In Section 5, we report the conducted experiments and the achieved results. Finally, Section 6 concludes the paper, presenting future directions.

2 RELATED WORK

Although most research in this area focuses on English, there are some initiatives in the Portuguese language. Here, we briefly present studies that focus solely on detecting toxic language.

Pelle et al. [26] developed the Hate2Vec model that relies on an ensemble-learning approach to classify offensive comments. The classifiers include (i) a lexicon-based classifier that leverages the semantic relatedness of word embeddings, (ii) a logistic regression classifier based on comment embeddings, and (iii) a standard Bag-of-Words (BoW) classifier based on unigram features. The authors evaluated their approach on datasets of the Portuguese and English languages, achieving an average F-score of 93%.

Soto et al. [35] explored distributional word embedding models with Convolutional Neural Networks (CNN) to detect hateful versus non-hateful comments. They applied 10-fold cross-validation to evaluate different dimensional configurations of pre-trained and fine-tuned Word2Vec and Wang2Vec embeddings [11] using the OffComBr and HLPHSD corpora [6, 9]. Their strategy achieved a performance of 92% F-score on HLPHSD and 86% on OffComBr.

Vargas et al. [39] approached a lexicon-based strategy to detect offensive language and Hate Speech. The authors utilized three lexical resources: Sentilex-PT [33], which provides semantic polarity, WorNetAffect.BR [24], which contains emotion types (e.g., anger, love, hate, disgust, and others), and an offensive contextual lexicon [37], which consists of explicit and implicit offensive and swearing expressions. They modeled these resources using a BoW structure and fed them into supervised machine learning algorithms. They evaluated the method on the HateBR Corpus [38], finding the best result of 88% of F-score for offensive language detection with Naïve Bayes, and 85% of F-score for Hate Speech detection with Multilayer Perceptron.

Saraiva et al. [31] modeled toxic comments over a heterogeneous network and addressed the detection of these comments through a semi-supervised algorithm applied to the network structure. They employed the LGC algorithm [43] to extract features from the network structure, which they then used to train several supervised machine learning algorithms. The approach was evaluated on the ToLD-BR corpus [16], achieving a 73% F-score.

Recent efforts have explored the use of APIs [14] and LLMs [21, 22] for toxicity detection. Kobellarz and Silva [14] employed the Perspective API [15] to evaluate toxicity in Brazilian Portuguese comments and their English translations. Their findings indicate that the API produces more reliable results when processing texts in the original language, suggesting that translation may distort toxicity detection. Moving beyond the API-based approach, Oliveira et al. [21] designed two prompt strategies, one with a narrow description and the other with more details, and then passed them to ChatGPT-3.5 Turbo [3] for toxic language detection in ToLD-BR [16] and HLPHSD [9] corpora. They report that the more detailed

prompt achieved the best performance, with an F-score of 74% for the HLPHSD corpus and an F-score of 73% for the ToLD-BR corpus.

In posterior research, Oliveira et al. [22] proposed an interactive prompt refinement process, in which top-rated prompts are iteratively adjusted until an optimal version is reached. Using a fine-tuned LLaMA 3.1 8B 4-bit model [10] on the ToLD-BR corpus, their method achieved competitive performance (75% F-score) compared to GPT-4o mini [23] and Sabiá 3 [1].

This work differs from previous studies because, in addition to addressing the binary task of toxicity detection, it also explores the more complex problem of detecting toxic language in a multi-label scenario, that is, identifying expressions with characteristics of racism, homophobia, xenophobia, among others. Our approach enhances the proposal by Saraiva et al. [31] by exploring an alternative method for regularizing heterogeneous graphs, while also employing contextual word embedding models that more accurately capture the semantic nuances inherent in toxic language. Like studies based on LLMs, our approach requires a smaller amount of labeled data for training, while also being more cost-efficient and computationally demanding.

3 CORPUS

To evaluate our strategy, we utilize the Toxic Language Dataset for Brazilian Portuguese (ToLD-Br) [16]. It has $21K$ tweets manually annotated into seven categories: LGBTQ+phobia, insult, xenophobia, misogyny, obscene, racism, and non-toxic. Additionally, this corpus has a binary version, which the authors released for use in binary classification tasks. Table 1 presents the distribution of tweets across each category and their respective proportions. As shown, the corpus exhibits a high degree of class imbalance. The labels with more tweets are non-toxic, obscene, and insult, while racism and xenophobia have fewer tweets.

Table 1: Distribution of tweets in the ToLD-Br corpus.

Label	Number	Proportion (%)	Type
LGBTQ+phobia	344	1.64	Multi-label
Insult	4,385	20.88	Multi-label
Xenophobia	151	0.72	Multi-label
Misogyny	463	2.20	Multi-label
Obscene	6,652	31.68	Multi-label
Racism	138	0.66	Multi-label
Non-toxic	8,867	42.22	Multi-label
Toxic	9,255	44.00	Binary
Non-toxic	11,745	56.00	Binary

In Table 2, we show examples of tweets exclusively belonging to the following labeled groups: insult, xenophobia, obscene, and racism, extracted from the corpus.

4 SEMI-SUPERVISED CLASSIFICATION METHOD

This section describes the techniques and word embeddings employed in conducting the experiments. As previously mentioned, we extend the strategy proposed by Saraiva et al. [31], a semi-supervised technique that classifies new instances using only a small amount of labeled data. The overall process follows a pipeline, as illustrated in Figure 2. The subsections 4.1, 4.2, 4.3, and 4.4 describe the stages.

Table 2: Examples of abusive language.

Label	Tweet
insult	*hoje é o caso de dormir cedo tipo um preso* (In English, "today is the case of sleeping early like a prisoner")
xenophobia	*crianças que dançam infinitamente mais que qualquer sulista* (In English "kids who dance infinitely more than any southerners")
obscene	*putz, e sex tape não dá pra gravar todo dia...* (In English "putz and sex tape can't be recorded every day...")
racism	*esse nego ta numa balaca nunca antes vista* (In English "this nigger is in an ostentation never before seen")

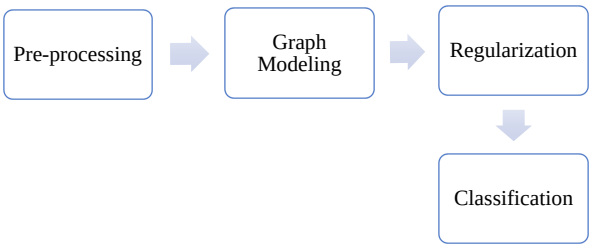

Figure 2: Pipeline for classifying toxic comments.

4.1 Pre-processing

We followed the same preprocessing procedures as proposed by [31]. Initially, we normalized abbreviated and repeated words using Enelvo [5], and then removed URLs, emojis, and emoticons to clean the tweets. Additionally, we extended their preprocessing by removing user mentions (e.g., `@user`) and retweet markers (e.g., `RT`).

4.2 Graph Modeling

Graphs are widely used structures for data representation and have gained increasing attention in the last decade. They have been successfully applied to a variety of tasks, such as topic modeling, name disambiguation, among others, often yielding promising results [13]. In particular, heterogeneous graphs encapsulate a large volume of information by combining structural relationships (edges) among nodes of different types with unstructured content associated with each node [42]. Their main contribution lies in making explicit the connections between distinct entities, enabling richer and more expressive modeling across various domains.

Building on the strategy proposed by Saraiva et al. [31], we model tweets as a heterogeneous graph composed of two node types: sentences and tokens. Edges are undirected and weighted, connecting only sentence nodes to token nodes, there are no sentence-to-sentence or token-to-token connections. While the original study relied on static word embeddings for edge weighting, we extend this approach by incorporating three contextual embedding models, two trained on Portuguese and one multilingual, with varying dimensionalities.

Formally, the heterogeneous graph is defined as $G = (V_t \cup V_s, E, W)$, where $V_t = \{v_{t1}, \ldots, v_{tn}\}$ is the set of token nodes, $V_s =$

$\{v_{s1}, \ldots, v_{sm}\}$ is the set of sentence nodes, $E = \{e_1, \ldots, e_k\}$ represents the set of edges, and W is the weighted adjacency matrix, where $W_{i,j}$ denotes the weight of the edge between nodes i and j. Each edge weight is computed as the average of the embedding vectors associated with the corresponding token node, as illustrated in Figure 3. Thus, each weight corresponds to a scalar value reflecting the token's embedding within its sentence context. Sentence nodes share token nodes whenever the same token appears across multiple sentences, allowing different sentences to interconnect through shared tokens.

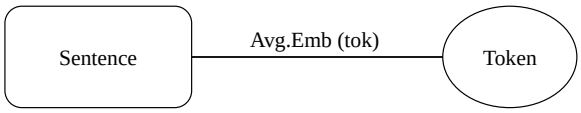

Figure 3: The network scheme for weighted edges.

To generate the embeddings, we employed three pre-trained contextual models designed for Portuguese and multilingual texts, applying them directly to the raw data without any fine-tuning or prompt engineering. Table 3 summarizes the models and the embedding dimensions considered. For the BERT-based models, we used the base versions, which produce 768-dimensional embeddings, while the LLaMA 3.1 8B model generates 4096-dimensional vectors. All models were implemented using the Hugging Face framework[1], and dimensionality reduction of the embeddings was performed using Principal Component Analysis (PCA) [17].

Table 3: Word embeddings and dimensions. ✓evaluated dimension and ✗not evaluated.

Model	Dimension				
	50	100	300	768	4,096
BERTimbau	✓	✓	✓	✓	✗
BERTweet.BR	✓	✓	✓	✓	✗
LLaMA 3.1 8B	✓	✓	✓	✓	✓

The first model, BERTimbau [36], is a Portuguese BERT model trained on 2.68 billion tokens from the brWaC corpus [40]. The second model, BERTweet.BR [4] follows the BERTweet-base [20] architecture and was trained from scratch on 100 million Portuguese tweets using the RoBERTa pre-training approach. The third, LLaMA 3.1 8B [10], is a multilingual model trained on 15 trillion tokens from public web sources.

Unlike static distributional embeddings such as Word2Vec [18] and GloVe [27], contextual models like BERT-based architectures and LLaMA 3.1 8B generate embeddings that more accurately capture the meanings of words within their specific contexts [34]. These embeddings generate unique vector representations for each word token by analyzing surrounding terms and the broader linguistic context in which the token appears.

[1]https://huggingface.co/

4.3 Regularization

The regularization is responsible for extracting features from the graph's objects. This method can be viewed as a form of transductive or semi-supervised classification that aims to determine a set of labels that satisfy two conditions: (i) consistency with the manually labeled data, and (ii) alignment with the graph topology, meaning that neighboring nodes are expected to share similar labels [29].

While Saraiva et al. [31] employed the Local and Global Consistency algorithm [43] during the regularization phase, our approach adopts the Gaussian Fields and Harmonic Functions algorithm [44]. Unlike LGC, our method preserves the original labels of the pre-labeled examples throughout the propagation process, ensuring consistency during label dissemination in low-data scenarios. In GFHF, the label of an unlabeled node is inferred by computing the weighted average of the label information from its neighboring nodes, where the weights correspond to the strength of the edges connecting them, as illustrated in Equation 1.

$$f_{o_i} = \frac{\sum\limits_{o_j \in O} w_{o_i,o_j} f_{o_j}}{\sum\limits_{o_j \in O} w_{o_i,o_j}} \tag{1}$$

The harmonic function is only applied to unlabeled objects, and the objective of the GFHF algorithm is to minimize the function in Equation 2.

$$\begin{aligned} Q(F) = \frac{1}{2} \sum_{o_i,o_j \in O} w_{o_i,o_j} (f_{o_i} - f_{o_j})^2 \\ + \lim_{u \to \infty} u \sum_{o_i \in O^L} (f_{o_i} - y_{o_i})^2 \end{aligned} \tag{2}$$

where:

- O is the set of nodes in the graph.
- O^L is the set of pre-annotated nodes in the graph.
- F is the regularization output. It represents a vector with the relative coordinates of a text in the plane.
- $\lim\limits_{u \to \infty} u$ does not allow the change of pre-annotated nodes.
- w is the weight of the edge between the nodes o_i and o_j.
- y is the information vector for the pre-annotated nodes.

To execute the GFHF algorithm, a set of pre-labeled nodes (O^L) must be provided to perform transductive classification. For instance, selecting 30% of pre-labeled nodes implies that 15% from each class will be randomly chosen for labeling. To infer the label of an unlabeled node, the algorithm computes the weighted average of its neighbors' label information, where the weights correspond to the edge strengths between nodes, as defined in Equation 1. Upon execution, the regularization algorithm produces coordinate values for each node in the graph, as illustrated in Table 4.

From Table 4, the **ID** is the object identifier, and the **Values** refers to the coordinates of each object in the network. **Label 1** shows toxicity (racism, xenophobia, insult, and so on), while **Label 0** is a non-toxic tweet.

4.4 Classification

In the final step, we fed a supervised machine learning algorithm with the output of the regularizer. We evaluated three classifiers: a

Table 4: Example of regularizer output.

ID	Value 1	Value 2	Label
355	0.007894	0.003272	1
467	0.001248	0.009521	0
749	0.003006	0.006123	0
133	0.006213	0.002590	1

Multi-Layer Perceptron (MLP)[2], a Support Vector Machine (SVM) using Stochastic Gradient Descent (SGDClassifier)[3], and Gradient Boosting (GB)[4], all from the Scikit-Learn library [25].

The implementation of our semi-supervised approach is available at https://github.com/fricarteneto/contextual-embeddings-semi-supervised-toxic-detection.

5 EXPERIMENTS AND RESULTS

This section describes the experiments and results performed for toxicity classification. In subsection 5.1, we detail our experiments in a binary classification setting, while in subsection 5.2, we present our evaluations in a multi-label classification problem.

5.1 Binary Classification

To evaluate the proposed method, we applied the supervised classifiers MLP, SVM, and GB to features extracted from the heterogeneous graph structure. To assess the contribution of contextual embeddings—used as edge weights in the graph—to toxic language detection, we experimented with both the GFHF regularization method and the LGC method employed in [31], while also investigating which proportion of pre-labeled data most effectively supports the detection of toxic content. For these experiments, the ToLD-BR dataset was split into 90% for training and 10% for testing. We tested pre-labeled data proportions of 10%, 20%, and 30% during the regularization step, using the embedding models in their original dimensions—768 for BERT-based embeddings and 4096 for LLaMA 3.1. Table 5 presents the macro F-score obtained by the three classifiers for each configuration across the evaluated proportions. The results suggest that regularization methods applied to contextual embeddings generally yield similar performance, even in low-resource scenarios with limited annotated data. Although both GFHF and LGC reached macro F-scores of up to 0.76, LGC exhibited greater performance variability across embeddings and classifiers—particularly for BERTweet.BR with the SVM classifier, which showed a substantial drop in performance. Still, both methods proved effective overall, with the highest macro F-score reaching 0.76, followed closely by 0.75, suggesting that GFHF and LGC yield comparable results in binary toxic language detection regardless of the proportion of pre-labeled data used.

We further explored the impact of contextual embeddings on the edge weighting between token and sentence nodes. Using 10% of pre-labeled data, the same proportion reported by Saraiva et al. [31], we evaluated contextual models with dimensions ranging from 50 to 768 for BERT-based embeddings and from 50 to 4096 for LLaMA 3.1. Table 6 presents the macro F-score results for each configuration. Overall, all embedding models delivered balanced and comparable results. The GB classifier demonstrated strong and consistent performance across all embedding models and dimensions, highlighting its robustness regardless of representation size. At 50 dimensions, both BERTweet.BR and LLaMA 3.1, when combined with the GB classifier, achieved a macro F-score of 0.76, outperforming all configurations based on BERTimbau across the tested dimensions. As the embedding dimensionality increased, both MLP and SVM classifiers improved their performance, with both achieving the second-highest scores (0.75) on BERTweet.BR and LLaMA 3.1 at their largest dimensions. At the highest dimensionality, all three classifiers with BERTimbau achieved competitive results; however, BERTweet.BR combined with GB stood out, achieving the best overall performance (0.76) and surpassing all other configurations. These findings suggest that the higher the embedding dimensionality, the better the model's ability to capture linguistic context relevant for toxicity detection. Notably, BERTweet.BR was pretrained on Portuguese tweets, the same domain as the ToLD-BR corpus, which likely enhances its capacity to capture the vocabulary and linguistic patterns commonly found on that platform.

Table 5: Macro F-score for binary toxic comment classification across different proportions of pre-labeled data, using GFHF and LGC regularization algorithms and contextual embedding models.

Alg.	Label (%)	BERTimbau			BERTweet.BR			LLaMA 3.1		
		MLP	SVM	GB	MLP	SVM	GB	MLP	SVM	GB
GFHF	10	0.74	0.75	0.75	0.75	0.75	**0.76**	0.75	0.75	0.75
	20	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	**0.76**
	30	0.75	0.75	0.75	**0.76**	0.75	**0.76**	0.75	0.75	**0.76**
LGC	10	0.74	0.74	0.75	0.40	0.74	**0.76**	0.74	0.74	**0.76**
	20	0.74	0.74	0.75	0.70	0.25	**0.76**	0.74	0.74	0.75
	30	0.74	0.74	0.75	0.70	0.25	**0.76**	0.74	0.74	**0.76**

Table 6: Macro F-score for binary toxic comment classification across different contextual embedding dimensions and models, using 10% of pre-labeled data with the GFHF regularization algorithm.

Dim.	BERTimbau			BERTweet.BR			LLaMA 3.1		
	MLP	SVM	GB	MLP	SVM	GB	MLP	SVM	GB
50	0.73	0.64	0.75	0.64	0.64	**0.76**	0.72	0.50	**0.76**
100	0.66	0.67	0.75	0.66	0.66	**0.75**	0.36	0.64	**0.75**
300	0.73	0.27	0.75	0.59	0.64	**0.76**	0.64	0.67	0.75
768	0.74	0.75	0.75	0.75	0.75	**0.76**	0.71	0.57	0.75
4096	–	–	–	–	–	–	**0.75**	**0.75**	**0.75**

Additionally, we compared the performance of our method with the semi-supervised approach proposed by Saraiva et al. [31], two

[2]Configured with 100 hidden units, ReLU activation function, Adam optimizer, learning rate of 0.001, and a maximum of 300 iterations.
[3]Employed with a linear kernel.
[4]Configured with 500 estimators and a maximum tree depth of 5.

transformer-based classifiers, monolingual (BR-BERT) and multilingual (M-BERT) [16], and finally, with the results reported from [21, 22] that employed ChatGPT-3.5 Turbo and LLaMA 3.1 for binary toxic detection. As shown in Table 7, our method outperformed the solutions based on LLMs [21, 22], the multilingual BERT classifier [16], and the semi-supervised approach [31]. In contrast, it achieved competitive performance relative to the monolingual transformer-based classifier [16]. It is worth emphasizing that our method attains this performance while relying on substantially fewer annotated samples (only 10%) than transformer-based methods, and with a lower computational cost compared to fine-tuned LLM-based solutions.

Table 7: Comparison between graph-based, transformer-based, and LLMs methods with our strategy.

Approach	Method	Macro F-score
Saraiva et al. [31]	GloVe + GB	0.73
Leite et al. [16]	BR-BERT	0.76
	M-BERT	0.75
Oliveira et al. [21]	ChatGPT-3.5	0.73
Oliveira et al. [22]	LLaMA 3.1	0.75
Ours	BERTweet.BR + GB	0.76

In summary, our results indicate that both the choice of embedding model and its dimensionality have a significant impact on classification performance. Gradient Boosting demonstrated consistency across all dimensions explored, while SVM and MLP achieved better performance at the highest dimensionality of the embedding models. All evaluated models demonstrated strong performance, with BERTweet.BR achieving the best overall results at 768 dimensions.

5.1.1 Error analysis. We performed an error analysis by examining classification error metrics and discussing representative examples of misclassification. In the context of toxic language detection, distinguishing between false positives and false negatives is particularly critical. False positives may lead to innocent individuals being banned from social networks or even facing legal consequences, while false negatives allow harmful content to continue circulating. Our analysis focused on the performance of classifiers using features derived from the graph structure with 768-dimensional BERTweet.BR embeddings, aiming to identify the primary sources of misclassification within the most effective configurations.

As illustrated in Figure 4, the classifiers produced relatively low rates of false negatives. Among the evaluated models, BERTweet.BR combined with SVM (Figure 4 (b)) achieved the lowest false negative rate. Representative examples are shown in Table 8. In Example 1, the classifiers failed to detect the derogatory connotation of the neologism "*petista*", a politically charged term commonly used with the intent to insult an individual or a group in Brazilian Portuguese. Correctly identifying such cases requires domain-specific knowledge. In Example 2, toxicity is conveyed subtly as the speaker criticizes the president's behavior, depicts him as lacking intelligence,

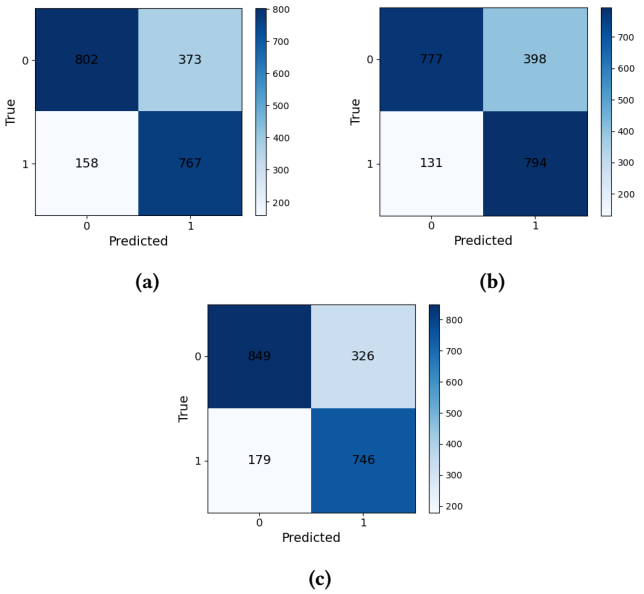

Figure 4: Confusion matrices of the classifiers evaluated using BERTweet.BR embeddings with 768 dimensions: (a) MLP; (b) SVM; (c) GB

and concludes with an ironic allusion to cannibalism. The indirect and sarcastic tone made detection challenging for all classifiers.

Table 8: Examples of false negatives identified by all three classifiers using 768-dimensional BERTweet.BR embeddings.

Example	Tweet
1	*ele provou chora petista* (In English, "he proved it, cry, petista")
2	*nunca vi na história do brasil um presidente tão desprovido de educação cadê o emprego presidente espero que não me mande comer gente* (In English, "I have never seen in the history of Brazil a president so devoid of manners, where is the job, president? I hope you do not tell me to eat people")

Despite the relatively low number of false negatives, classifier performance was strongly affected by a high incidence of false positives. Among the evaluated configurations, BERTweet.BR combined with GB (Figure 4 (c)) exhibited the lowest false positive rate. These cases are primarily triggered by pejorative or negatively connoted terms, such as "*desgraçada*" (wretch), "*burro*" (donkey), and "*puta*" (whore, used here to mean "angry"), which biases the models toward labeling sentences as toxic even when they occur in neutral or humorous contexts. Examples of such tweets with spurious correlations are shown in Table 9.

These results suggest that, even with the use of contextual embedding models within this semi-supervised framework, which relies on a limited amount of labeled data, the classifiers still struggled to detect implicit toxicity or domain-specific cases, such as politically charged language, which often require external knowledge to be properly identified. Another limiting factor is the presence of spurious lexical cues, which inflate the false positive rate by causing the models to over-rely on isolated terms rather than contextual meaning. These challenges highlight the need to investigate approaches

Table 9: Examples of false positives identified by all three classifiers using 768-dimensional BERTweet.BR embeddings.

Example	Tweet
3	*chorando feito uma desgraçada* (In English, "crying like a wretch")
4	*esse é o burro do shrek* (In English, "this is Shrek's donkey")
5	*minha mãe está estão puta comigo mds* (In English, "my mom is mad at me omg")

for bias mitigation [19] and domain-aware [12, 30], which could enhance the robustness of this semi-supervised framework.

5.2 Multi-label Classification

We transformed the multi-label task learning into six independent binary classification problems to predict the six toxic groups of the ToLD-BR corpus. For this task, we followed the same data splitting strategy used in the binary classification experiments, dividing each binary problem into 90% for training and 10% for testing. Unlike the binary classification setting, where only 10% of pre-labeled data was used for the GFHF algorithm, we increased the proportion of pre-labeled data to 30% in this stage due to the limited number of comments in certain toxic categories (e.g., xenophobia and racism)[5]. The same classification algorithms used in the binary toxicity classification task were tested in the multi-label setting. However, we only report the results for the GB algorithm, as it significantly outperformed the others.

Table 10 summarizes the average Precision and *Hamming* loss of the toxic groups for each embedding model, along with its respective dimension size. As in the binary classification task, the average precision improves as the dimensionality of the embeddings increases. The best performance was achieved by the LLaMA 3.1 model, which reaches an average precision of 0.30 and a Hamming loss of 0.07 at 4096 dimensions. Among the BERT-based models, BERTimbau performs best at the lowest dimension, while BERTweet.BR consistently outperforms it in all other dimensions, achieving its highest precision (0.28) at the 768-dimensional representation.

Table 10: Average precision and *Hamming* loss for multi-label toxic comment classification across different contextual embedding dimensions and models, using 30% of pre-labeled data, the GFHF regularization algorithm, and the Gradient Boosting classifier.

Dimension	Classifier	Models					
		BERTimbau		BERTweet.BR		LLaMA 3.1	
		Precision	*Hamming*	Precision	*Hamming*	Precision	*Hamming*
50		**0.27**	**0.08**	0.26	0.07	**0.27**	**0.08**
100		**0.27**	**0.08**	**0.27**	**0.08**	**0.27**	**0.08**
300	**Gradient Boosting**	0.26	0.07	**0.27**	**0.08**	0.26	0.08
768		0.27	0.08	**0.28**	**0.07**	0.28	0.08
4096		–	–	–	–	**0.30**	**0.07**

We compile the best performances achieved by contextual models in Table 11, along with a comparison against the experiment conducted by Leite et al. [16] using the same dataset. The authors applied the same binary relevance method to transform the multi-label task learning into independent binary classification problems, using the same proportion of 90% - 10% for training and testing, respectively. To classify toxic comments, they used a Bag-of-Words + AutoML [8] model and a multilingual BERT [7] model. It can be observed that our models outperformed both approaches by Leite et al. [16] in terms of average precision, using only 30% of pre-labeled data. Nonetheless, our *Hamming* Loss was slightly higher compared to theirs.

Table 11: Comparison of average precision and *Hamming* loss results between our method versus BoW-based classifier, and multilingual BERT.

Model	Classifier	Avg. precision	*Hamming* loss
BERTimbau		0.27	0.08
BERTweet	**Gradient Boosting**	0.28	0.07
LLaMA 3.1		**0.30**	**0.07**
-	BoW + AutoML [16]	0.20	0.08
	M-BERT [16]	0.19	**0.07**

In addition to the average precision and *Hamming* loss results, we analyzed the F-score metric for each toxic group, comparing the results with both methods presented by Leite et al. [16], as shown in Table 12. According to the results, the LLaMA 3.1 + GB model outperforms both BoW + AutoML and M-BERT in terms of the average F-score across toxic labels. Our approach also outperforms both methods individually on each label, except for the *insult* category, where M-BERT achieves slightly better performance. It is important to note that for underrepresented labels such as LGBT+phobia, racism, and xenophobia, our model achieved superior performance, correctly identifying tweets from these categories despite their low frequency.

Table 12: F-score results for each toxic group.

Label	F-score		
	LLaMA 3.1 + GB	BoW + AutoML	M-BERT
LGBT+phobia	**0.55**	0.34	0.07
Obscene	**0.65**	0.63	0.46
Insult	0.44	0.37	**0.45**
Racism	**0.32**	0.00	0.00
Misogyny	**0.47**	0.34	0.17
Xenophobia	**0.33**	0.00	0.00
Average	**0.46**	0.28	0.38

5.2.1 Error Analysis. We also performed an error analysis for multi-label classification task. In Tables 13 and 14, we show the results for false negatives and false positives, respectively, highlighting the lowest values for each metric. It can be observed that the LLaMA 3.1 + GB model produces fewer false negatives than BoW + AutoML and M-BERT, except in the *insult* category, where M-BERT performed slightly better. This indicates that our approach is more effective

[5]Experiments with 10% and 20% of pre-labeled data were also conducted; however, the semi-supervised method failed to detect some minority toxic classes under those settings.

at preventing abusive content from remaining undetected. On the other hand, as shown in Table 14, the LLaMA 3.1 + GB model is outperformed by both BoW + AutoML and M-BERT, which achieve the lowest false positive rates, suggesting they are more precise than LLaMA 3.1 + GB in avoiding incorrect positive classifications.

Table 13: False negative metric for the classification task.

Label	False Negative		
	LLaMA 3.1 + GB	BoW + AutoML	M-BERT
LGBT+phobia	**17**	20	25
Obscene	**207**	263	427
Insult	295	322	**290**
Racism	**11**	**11**	**11**
Misogyny	**29**	34	39
Xenophobia	**12**	19	19

Table 14: False positive metric for the classification task.

Label	False Positive		
	LLaMA 3.1 + GB	BoW + AutoML	M-BERT
LGBT+phobia	11	3	**2**
Obscene	278	167	**38**
Insult	75	28	**41**
Racism	2	**0**	**0**
Misogyny	9	1	**0**
Xenophobia	**0**	**0**	**0**

An analysis of misclassified tweets from our approach in the multi-label classification task shows that, similar to the binary setting, it struggles to accurately identify toxicity when irony is present. For example, the homophobic tweet "essas meninas com pau são as mais lindas" ("these girls with a penis are the most beautiful") was not correctly flagged as toxic. Moreover, our approach produced false negatives even in tweets with explicitly abusive language, such as the insult-labeled example "Lixo, esgoto, fossa a céu aberto, tenho nojo de chamar um ser como você de humano" ("Trash, sewage, open sewer, I am disgusted to call a being like you human"). In the case of false positives, spurious terms once again played a significant role in the multi-label classification context. The tweet "amiga perdi 50% da minha postura de piranha" ("friend, I lost 50% of my slutty posture") was incorrectly classified as misogynistic, while "pensa numa bixa realizada" ("think of a fulfilled fag") was wrongly labeled as homophobic due to the presence of the expression "bixa realizada" (fulfilled fag).

Overall, the results indicate that contextual embeddings capture nuanced linguistic context more effectively, which is essential for detecting toxic language within heterogeneous graph structures. Furthermore, increasing embedding dimensionality generally enhances model performance. These findings underscore the importance of leveraging contextualized representations when modeling complex language phenomena, such as toxicity, particularly when integrated with graph-based learning frameworks. However, as in the binary classification task, the semi-supervised approach failed to correctly identify toxicity in tweets containing irony. Additionally, negatively connoted expressions used out of context often led to false positives, highlighting the challenge posed by abusive terms that bias the approach.

6 CONCLUSION

In this paper, we investigated the use of contextual word embedding models within a semi-supervised framework based on heterogeneous graphs to address both binary and multi-label toxic comment classification. Our findings demonstrate that the choice of embedding model and its dimensionality significantly influence classification performance. Using only 10% of labeled data, the proposed approach outperformed ChatGPT-3.5 Turbo and fine-tuned LLaMA 3.1 [21, 22], while remaining competitive with transformer-based solutions [16], achieving an F-score of 0.76 in the binary classification task. Furthermore, the semi-supervised graph-based approach using edge weights derived from contextual embeddings yielded slightly better results than its counterpart based on static distributional models [31], indicating that contextual embeddings are more effective at encoding toxic language. In the multi-label scenario, high-dimensional embeddings proved particularly beneficial, with the configuration with 4096-dimensional LLaMA 3.1 achieving the best performance: an average F-score of 0.46, precision of 0.30, and Hamming loss of 0.07. Moreover, our approach outperformed the methods reported by Leite et al. [16] across multiple metrics, despite using only 30% of the training data. Nevertheless, the semi-supervised method exhibited limitations, such as failing to reliably detect toxicity in tweets containing irony and producing false positives due to negatively connoted terms used out of context. These challenges highlight the need for debiasing strategies and mechanisms capable of capturing subtle cues such as sarcasm and irony. In future work, we intend to explore bias mitigation [19] and domain-aware techniques [12, 30], as well as the integration of offensive lexicons to enhance the robustness of this semi-supervised framework. Furthermore, we plan to experiment with contextual word embedding models derived from more robust LLMs to weight the edges of the heterogeneous graph, particularly the SoberanIA model[6], developed by the Secretaria de Inteligência Artificial, Economia Digital, Ciência, Tecnologia e Inovação do Piauí (SIA/PI) using public and governmental data. We expect that embeddings generated by more robust LLMs will capture richer contextual information, enhancing the encoding of linguistic nuances and, consequently, improving toxicity detection. We also aim to investigate alternative transformation methods for multi-label learning and analyze complex network metrics to better understand relationships among toxic labels.

ACKNOWLEDGMENTS

The authors acknowledge the support of the SoberanIA Project, Governo do Estado do Piauí (SIA/PI and PIT/ETIPI), IFPI, and UFPI (DCCMAPI) for enabling the development of this research.

[6] https://soberania.ai/

REFERENCES

[1] Hugo Abonizio, Thales Sales Almeida, Thiago Laitz, Roseval Malaquias Junior, Giovana Kerche Bonás, Rodrigo Nogueira, and Ramon Pires. 2025. Sabiá-3 Technical Report. arXiv:2410.12049 [cs.CL]

[2] Gabriel Assis, Annie Amorim, Jonnathan Carvalho, Daniel de Oliveira, Daniela Vianna, and Aline Paes. 2024. Exploring Portuguese Hate Speech Detection in Low-Resource Settings: Lightly Tuning Encoder Models or In-Context Learning of Large Models?. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, Pablo Gamallo, Daniela Claro, António Teixeira, Livy Real, Marcos Garcia, Hugo Gonçalo Oliveira, and Raquel Amaro (Eds.). Association for Computational Lingustics, Santiago de Compostela, Galicia/Spain, 301–311.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and et al. Askell. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., Online, 1877–1901.

[4] Fernando Carneiro, Daniela Vianna, Jonnathan Carvalho, Alexandre Plastino, and Aline Paes. 2024. BERTweet.BR: a pre-trained language model for tweets in Portuguese. *Neural Computing and Applications* 37 (12 2024), 4363–4385.

[5] Thales Felipe Costa Bertaglia and Maria das Graças Volpe Nunes. 2016. Exploring Word Embeddings for Unsupervised Textual User-Generated Content Normalization. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*. The COLING 2016 Organizing Committee, Osaka, Japan, 112–120.

[6] Rogers Prates de Pelle and Viviane P Moreira. 2017. Offensive Comments in the Brazilian Web: a dataset and baseline results. In *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*. Sociedade Brasileira de Computação, São Paulo, Brazil, 510–519.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186.

[8] Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Springenberg, Manuel Blum, and Frank Hutter. 2015. Efficient and Robust Automated Machine Learning. In *Proceedings of the Twenty-ninth Conference on Neural Information Processing Systems*. Curran Associates, Inc., Montreal, Canada.

[9] 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*. Association for Computational Linguistics, Florence, Italy, 94–104.

[10] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, and et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783 [cs.AI]

[11] Nathan Hartmann, Erick Fonseca, Christopher Shulby, Marcos Treviso, Jéssica Silva, and Sandra Aluísio. 2017. Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks. In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*. Sociedade Brasileira de Computação, Uberlândia, Brazil, 122–131.

[12] Aiqi Jiang and Arkaitz Zubiaga. 2023. SexWEs: Domain-Aware Word Embeddings via Cross-Lingual Semantic Specialisation for Chinese Sexism Detection in Social Media. *Proceedings of the International AAAI Conference on Web and Social Media* 17, 1 (Jun. 2023), 447–458. https://doi.org/10.1609/icwsm.v17i1.22159

[13] Ben King, Rahul Jha, and Dragomir R. Radev. 2014. Heterogeneous Networks and Their Applications: Scientometrics, Name Disambiguation, and Topic Modeling. *Transactions of the Association for Computational Linguistics* 2 (2014), 1–14.

[14] Jordan K. Kobellarz and Thiago H. Silva. 2022. Should We Translate? Evaluating Toxicity in Online Comments when Translating from Portuguese to English. In *Proceedings of the Brazilian Symposium on Multimedia and the Web* (Curitiba, Brazil) *(WebMedia '22)*. Association for Computing Machinery, New York, NY, USA, 89–98. https://doi.org/10.1145/3539637.3556892

[15] Alyssa Lees, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 2022. A New Generation of Perspective API: Efficient Multilingual Character-level Transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Washington DC, USA) *(KDD '22)*. Association for Computing Machinery, New York, NY, USA, 3197–3207. https://doi.org/10.1145/3534678.3539147

[16] 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*. Association for Computational Linguistics, Suzhou, China, 914–924.

[17] Andrzej Maćkiewicz and Waldemar Ratajczak. 1993. Principal components analysis (PCA). *Computers & Geosciences* 19, 3 (1993), 303–342. https://doi.org/10.1016/0098-3004(93)90090-R

[18] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations, Workshop Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). Scottsdale, Arizona, USA.

[19] Francimaria R.S. Nascimento, George D.C. Cavalcanti, and Márjory Da Costa-Abreu. 2022. 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), 117032. https://doi.org/10.1016/j.eswa.2022.117032

[20] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Qun Liu and David Schlangen (Eds.). Association for Computational Linguistics, Online, 9–14.

[21] 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* (Belo Horizonte/MG). SBC, Porto Alegre, RS, Brasil, 94–103.

[22] Amanda Oliveira, Pedro H. Silva, Valéria Santos, Gladston Moreira, Vander L. Freitas, and Eduardo J. Luz. 2024. Toxic Text Classification in Portuguese: Is LLaMA 3.1 8B All You Need?. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana* (Belém/PA). SBC, Porto Alegre, RS, Brasil, 57–66.

[23] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, and et al. Janko Altenschmidt. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[24] Paulo Roberto Pasqualotti and Renata Vieira. 2008. WordnetAffectBR: uma base lexical de palavras de emoções para a língua portuguesa. *RENOTE* 6, 1 (2008).

[25] 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 (2011), 2825–2830.

[26] 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*. Association for Computing Machinery, Salvador, BA, Brazil, 237–243.

[27] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1532–1543.

[28] Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2020. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation* 55 (2020), 477–523.

[29] Rafael Geraldeli Rossi. 2015. *Classificação automática de textos por meio de aprendizado de máquina baseado em redes*. Ph. D. Dissertation. Instituto de Ciências Matemáticas e de Computação. http://www.teses.usp.br/teses/disponiveis/55/55134/tde-05042016-105648

[30] Parisa Safikhani and David Broneske. 2025. AutoML Meets Hugging Face: Domain-Aware Pretrained Model Selection for Text Classification. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, Abteen Ebrahimi, Samar Haider, Emmy Liu, Sammar Haider, Maria Leonor Pacheco, and Shira Wein (Eds.). Association for Computational Linguistics, Albuquerque, USA, 466–473. https://doi.org/10.18653/v1/2025.naacl-srw.45

[31] Ghivvago D Saraiva, Rafael T Anchiêta, Franciso A R Neto, and Raimundo S Moura. 2021. A Semi-Supervised Approach to Detect Toxic Comments. In *Proceedings of the International Conference on Recents Advances in Natural Language Processing*. INCOMA Ltd., Online, 1265–1271.

[32] Chuan. Shi and Philip S. Yu. 2017. *Heterogeneous Information Network Analysis and Applications* (1st ed. ed.). Springer International Publishing, Cham.

[33] Mário J Silva, Paula Carvalho, and Luís Sarmento. 2012. Building a sentiment lexicon for social judgement mining. In *Proceedings of the 10th International Conference on Computational Processing of the Portuguese Language*. Springer, Coimbra, Portugal, 218–228.

[34] Noah A. Smith. 2020. Contextual word representations: putting words into computers. *Commun. ACM* 63, 6 (May 2020), 66–74.

[35] Claver Soto, Gustavo Nunes, and José Gomes. 2019. Avaliação de técnicas de word embedding na tarefa de detecção de discurso de ódio. In *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional* (Salvador). SBC, Porto Alegre, RS, Brasil, 1020–1031.

[36] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Intelligent Systems*, Ricardo Cerri and Ronaldo C. Prati (Eds.). Springer International Publishing, Cham, 403–417.

[37] Francielle Vargas, Isabelle Carvalho, Thiago A. S. Pardo, and Fabrício Benevenuto. 2024. Context-aware and expert data resources for Brazilian Portuguese hate speech detection. *Natural Language Processing* (2024), 1–22. https://doi.org/10.1017/nlp.2024.18

[38] 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*. European Language Resources Association, Marseille, France, 7174–7183.

[39] Francielle Vargas, Fabiana Goés, Isabelle Carvalho, Frabrício Benevenuto, and Thiago A S Pardo. 2021. Contextual-Lexicon Approach for Abusive Language Detection. In *Proceedings of the International Conference on Recents Advances in Natural Language Processing*. INCOMA Ltd., Online, 1438–1447.

[40] Jorge A. Wagner Filho, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. 2018. The brWaC Corpus: A New Open Resource for Brazilian Portuguese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*. European Language Resources Association, Miyazaki, Japan, 4339–4344.

[41] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Minneapolis, Minnesota, USA, 75–86.

[42] Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V. Chawla. 2019. Heterogeneous Graph Neural Network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Anchorage, AK, USA). Association for Computing Machinery, New York, NY, USA, 793–803.

[43] Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. 2004. Learning with local and global consistency. In *Advances in neural information processing systems*. MIT Press, MA, USA, 321–328.

[44] Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. 2003. Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*. AAAI Press, Washington, DC, USA, 912–919.