A Semi-Supervised Approach to Detect Toxic Comments

Ghivvago D. Saraiva[1], Rafael T. Anchiêta[2], Francisco A. R. Neto[2] and Raimundo S. Moura[1]

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

Abstract

Toxic comments contain forms of non-acceptable language targeted towards groups or individuals. These types of comments become a serious concern for government organizations, online communities, and social media platforms. Although there are some approaches to handle non-acceptable language, most of them focus on supervised learning and the English language. In this paper, we deal with toxic comment detection as a semi-supervised strategy over a heterogeneous graph. We evaluate the approach on a toxic dataset of the Portuguese language, outperforming several graph-based methods and achieving competitive results compared to transformer architectures.

1 Introduction

Toxic comments, posts, and other types of content became more common in social media nowadays. They contain forms of non-acceptable language (profanity), which may be concealed or explicit, including insults and threats directed to a group or individual (Zampieri et al., 2019). These comments spread rapidly on the internet, especially on social networks where they find acceptance, and may culminate in several threats to individuals, becoming a serious concern for government organizations, online communities, and social media platforms.

The term toxic comment is commonly found in literature as harmful speech, hate speech, or offensive language. Toxic comment may be viewed as negative online behaviors, i.e., comments that are rude, disrespectful, may contain hate speech, or otherwise likely to make someone leave a discussion[1]. Schmidt and Wiegand (2017) define hate speech as any communication that disparages a person or a group based on some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics. Also, it may occur with different linguistic styles, even in subtle forms or when humour is used (Fortuna and Nunes, 2018). It is important to highlight that fighting these types of comments is of utmost importance since they are a crime in several countries.

To deal with toxic comments, most approaches adopt supervised-machine learning techniques and are mainly focused on the English language (Poletto et al., 2020). These approaches range from surface-level features, as Bag-Of-Words (Paiva et al., 2019), linguistics features, as Part-Of-Speech information (Chen et al., 2012), deep neural networks, as Long Short-Term Memory (LSTM) (Fortuna et al., 2019) and Convolutional Neural Networks (CNN) (Badjatiya et al., 2017) to Transformer architectures (Leite et al., 2020). Despite interesting results achieved by Transformer architectures, there are still several rooms to be explored in this research area.

In this paper, we developed a semi-supervised strategy to detect toxic comments in the Brazilian Portuguese language. Semi-supervision is the problem of learning from labeled and unlabeled data (Abney, 2007; Subramanya and Talukdar, 2014), in which given a point set $X = \{x_1, ..., x_l, x_{l+1}, ..., x_n\}$ and a label set $L = \{1, ..., c\}$, the first l points have labels $\{y_1, ..., y_l\} \in L$ and the remaining points are unlabeled (Zhou et al., 2004).

We modeled that problem as a heterogeneous network. The structure of our graph was inspired by de Sousa et al. (2020) and Anchiêta et al. (2020). These authors modeled the tasks of helpfulness prediction and paraphrase identification as a heterogeneous network, respectively. For that, they defined an undirected unweighted graph with two node types: sentence and token. However, we have

[1] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/overview

Proceedings of Recent Advances in Natural Language Processing, pages 1261–1267
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_142

created a weighted graph based on pre-trained word embeddings. The weight between sentence and token nodes is the average of the embedding values for that token. Figure 1 depicts an example of a sentence modeled as a graph. From this figure, we may see two node types: token and sentence, and an undirected and weighted edges between the sentence and tokens nodes.

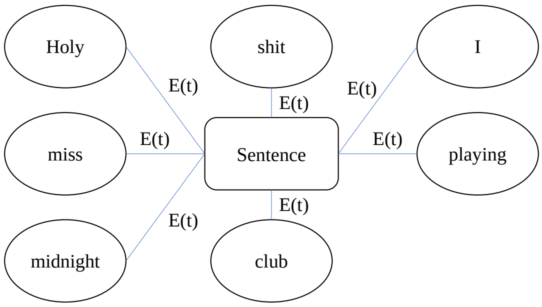

Figure 1: Example of a graph model for the sentence "*Holy shit, I miss playing midnight club*".

To extract features from the graph structure, we used a regularization algorithm that propagates labels from a small set of labeled nodes to the entire graph.

We evaluated the approach using the `ToLD-Br` corpus (Leite et al., 2020). It has twenty-one thousand annotated tweets as either toxic or non-toxic language. Also, we compared our strategy with different graph-based methods and with transformer-based methods. Our method outperformed all graph-based approaches and achieved competitive results compared to transformer-based methods, using only 10% of labeled nodes.

The reminder of this paper is structured as follows: Section 2 briefly presents related work. In Section 3, we show the used corpora. Section 4 details our developed approach. In Section 5, we analyze the conducted experiments. Finally, Section 6 concludes the paper, presenting future directions.

2 Related Word

As aforementioned, the main approaches to detect toxic comments are based on supervised machine learning. Here, we briefly present the main works.

Most of the works that study this task commonly point first to surface-level features, such as bag of words and lexicon-based approaches, with negative words as features (Gitari et al., 2015; Waseem and Hovy, 2016; Waseem et al., 2017; Schmidt and Wiegand, 2017).

More recently, neural networks-based strategies and transformer-based architectures has been applied to hate speech detection due to the good results achieved in various tasks. Banerjee et al. (2020) evaluated pre-trained word embeddings with CNN networks to hate speech detection for the Indian language. Rizwan et al. (2020) explored transfer-learning of embeddings models to Roman Urdu and developed a CNN-gram network to hate speech classification for that language. Duwairi et al. (2021) investigated the ability of CNN, CNN-LSTM, and BiLSTM-CNN to classify hate speech in Arabic. Plaza-del Arco et al. (2021) compared two pre-trained language models, such as BERT (Devlin et al., 2019) and XLM (CONNEAU and Lample, 2019) trained to detect hate speech in the Spanish language.

For the Portuguese language, most of the works follow the trend of supervised approaches. de Pelle and Moreira (2017) created a dataset consist of $1,250$ offensive comments and developed a baseline method based on n-gram features to classify offensive comments in their dataset. Fortuna et al. (2019) created a hate speech dataset composed of $5,668$ tweets and developed a baseline classification using pre-trained word embeddings and LSTM in their dataset. Coutinho and Malheiros (2020) trained a logistic regression using superficial features for sentiment analysis. Then, they evaluated that model into a homophobia corpus to detect homophobic posts.

Although there are some efforts to detect non-acceptable language in Portuguese, they evaluate the developed approach in their own corpus, making a fair comparison among the models difficult. Moreover, these corpora are much smaller when compared to corpora of other languages (Poletto et al., 2020) and than the `ToLD-Br` corpus. This fact makes the development of robust strategies to handle toxic comments difficult, as they usually require a large corpus.

3 ToLD-BR Corpus

Toxic Language Dataset for Brazilian Portuguese (ToLD-Br) (Leite et al., 2020) is a very recent dataset with Twitter posts in the Brazilian Portuguese language. It has $21K$ tweets manually annotated into seven categories: *non-toxic*, *LGBTQ+phobia*, *obscene*, *insult*, *racism*, *misogyny*, and *xenophobia*. The corpus is the largest dataset available for toxic data analysis in social media for Portuguese and the first dataset with demographic information about annotators.

Besides seven categories, the authors released a binary version of the corpus for the binary classification task, as shown in Table 1.

Label	Train.	Valid.	Test	Prop.
Toxic	7,375	908	972	44%
Non-toxic	9,425	1,192	1,128	56%

Table 1: Binary version of the ToLD-Br corpus.

As one can see in Table 1, the corpus has a little more non-toxic than toxic tweets. In this paper, we adopted the binary version of the corpus, i.e., our objective is to identify if a comment is toxic or non-toxic.

In what follows, we detail our strategy to handle toxic texts.

4 Semi-supervised approach

We organized the strategy into four steps, as illustrated in Figure 2. Subsections 4.1, 4.2, 4.3, and 4.4 describe the stages.

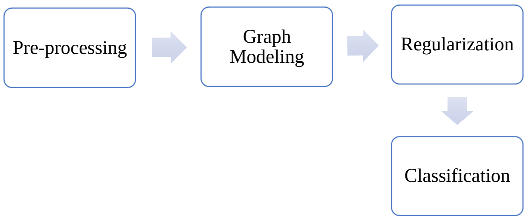

Figure 2: Process to deal with toxic comments.

4.1 Pre-processing

In the pre-processing[2], we normalized and cleaned the tweets. In the first one, we applied the Enelvo tool (Costa Bertaglia and Volpe Nunes, 2016) to normalise abbreviated and repeated words. In the second one, we simply clean URLs, emojis, and tweet mentions.

4.2 Graph-Based Method

We modeled toxic comments detection as a heterogeneous network since this network type contains abundant information with structural relations (edges) among multi-typed nodes as well as unstructured content associated with each node (Zhang et al., 2019). Graph structures have been used for several tasks, such as: topic model, name disambiguation, scientific impact measurement, and others, obtaining good results (King et al., 2014).

We defined a undirected and weighted graph as $G = (V, E, W)$, where V is a set of vertices $V = \{v_1, ..., v_n\}$, E indicates a set of edges $E = \{e_1, ..., e_n\}$, and W is a weighted adjacency matrix, in which $W_{i,j}$ denotes the weight of an edge between nodes i and j. We defined two node types: token and sentence and two constraints not allowing link among tokens nodes or among sentences nodes.

The strategy of weighting links between a token and a sentence node is straightforward. The weight is the average[3] of embedding vectors of the token node. To get embedding values for each token, we used 100-dimensional GloVe embeddings[4] for the Portuguese language (Hartmann et al., 2017). Figure 3 shows the scheme of the network designed for this task.

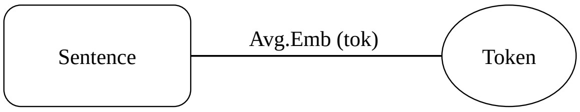

Figure 3: The network scheme for weighted edges.

One can see that the edges are undirected and weighted, and a sentence node may share several token nodes whenever the token is in the sentence, i.e., the edges between token nodes and sentence nodes are based on word occurrence in sentence.

4.3 Regularization

To extract the features regarding the network object classes, we applied a regularization method to the graph. Regularization is a kind of semi-supervised (or transductive) classification method that aims to find a set of labels, minimizing a cost function and satisfying two conditions: (i) the method needs to be consistent with the set of labels manually annotated and (ii) the method needs to be consistent with the network topology, considering that nearest neighbors tend to have the same labels (Ji et al., 2010).

We used the learning with Local and Global Consistence (LGC) (Zhou et al., 2004) as a regularization method. The algorithm designs a classi-

[2] The obtained results without pre-processing were worse than with pre-preprocessing.

[3] We also tested the sum, maximum, and minimun values.

[4] We also experimented other pre-trained models with dimensions of 50, 100, and 300.

fying function that is sufficiently smooth concerning the intrinsic structure collectively revealed by known labeled and unlabeled points. Thus, the LGC lets every point iteratively spread its label information to its neighbors until a global stable state is achieved (Gui et al., 2014). Also, it allows the class information of the labeled objects to be changed during the classification as objects may be erroneously labeled and, consequently, decrease the performance of the classification. More than that, the algorithm diminished the influence of objects with a high degree (many neighboring objects), therefore, these objects do not have excessive influence in the classification.

To execute the algorithm, a set of nodes need to be pre-labeled. The regularizer randomly pre-labeled, i.e., supposing that the percentage of pre-labeled nodes is equals 5%, it means that 0.25% of each class is randomly pre-labeled. As a result, the regularizer produces values related to coordinates for each object in the network, as shown in Table 2.

Id	**Value 1**	**Value 2**	**Label**
100	0.004567	0.001456	1
255	0.002789	0.008763	0
878	0.001998	0.005342	0
233	0.008764	0.003215	1

Table 2: Example of regularizer output.

From Table 2, **Id** is the object identifier, **Values** refer to coordinates of each object in the network, and **Label 1** shows toxic, while **Label 0** is a non-toxic tweet.

4.4 Classification

With the regularization values, we fed several machine learning algorithms to identify and predict toxic comments. We experimented Multi Layer Perceptron, Naïve Bayes, Decision Tree, Support Vector Machine, and Gradient Boosting from the Scikit-Learn library (Pedregosa et al., 2011).

In the following section, we detailed our carried out experiments, then, the achieved results are presented.

5 Experiments and Results

In order to produce coordinate values for each object from the regularizer, we ranged the number of pre-labeled nodes from 5% to 30%. Then, we applied the machine learning algorithms to train and classifier toxic comments.

We achieved the best result with the Gradient Boosting classifier[5] using only 10% of the pre-labeled nodes i.e., the classification does not improve after this percentage. Table 3 shows the achieved results. It is important to say that only the training set is pre-labeled.

Pre-labeled (%)	**F-score**	
	Toxic	**Non-toxic**
5	0.73	0.73
10	0.73	0.74
...	...	...

Table 3: Achieved results with the gradient boosting classifier.

Besides our approach, we evaluated other graph models of different structures. First, we used the network graph developed by Anchiêta et al. (2020). That graph does not use weight between the nodes. Second, we used the Term Frequency-Inverse Document Frequency (TF-IDF) as weight instead of the average of embeddings. Third, we used bigrams and trigrams as nodes rather than token nodes. Finally, we used the Pointwise Mutual Information (PMI) measure (Church and Hanks, 1990) as the weight between the bi and trigrams nodes. For these approaches, we adopted the same regularization algorithm, ranging the pre-labeled nodes from 5% to 30%. In Table 4, we present the best-achieved results.

Pre-labeled	**Method**	**F-score**		**Classifier**
		Toxic	**Non-toxic**	
30%	Trigrams without weight	0.70	0.40	MLP
30%	Bigrams without weight	0.72	0.53	MLP
30%	Trigrams + PMI	0.69	0.51	GB
30%	Bigrams + PMI	0.62	0.39	GB
30%	Unigrams + TF-IDF	0.69	0.62	GB
30%	Anchiêta et al. (2020)	0.68	0.61	MLP

Table 4: Comparison among graph-based approaches.

From this table, our graph modeling and the gradient boosting classifier achieved better results than these other graphs, as well as classifier variations. This, we think, is because of the embedding value among the graph nodes since it is able to capture morphological, syntactic, and semantic knowledge of a word. As we used the average word embedding value, it includes information from all of the

[5] We used as parameters $n\_stimators = 5$ and $max\_depth = 5$

individual vector values, working as an overall summary of all vector values.

We further compared our strategy with other graph-based approaches: Text Graph Convolutional Network (TextGCN) (Yao et al., 2019) and Heterogeneous Graph Attention Network (HGAT) (Yang et al., 2021). The former models the whole text corpus as a document-word graph with word co-occurrence relations and applies GCN for classification. The latter models the texts using a heterogeneous information network framework and adopts heterogeneous graph attention to embed that framework for text classification based on a dual-level attention mechanism. Finally, we compared our approach with a transformer-based method as it has achieved remarkable results in several areas of Natural Language Processing (NLP). We compared our strategy with BR-BERT (Leite et al., 2020), which is a monolingual BERT, and M-BERT (Leite et al., 2020), which is a multilingual BERT. Table 5 shows the comparison between these methods.

Approach	Model	F-score		Macro F-score
		Toxic	Non-toxic	
Graph	TextGCN	0.70	0.69	0.69
	HGAT	0.55	0.50	0.53
Transformer	BR-BERT	0.79	0.74	0.76
	M-BERT	0.77	0.75	0.75
Graph	Ours	0.73	0.74	0.73

Table 5: Comparison between graph-based and transformer-based methods with our strategy.

As we can see from Table 5, our approach outperformed the graph-based methods and reached a competitive result compared to transformer models. Although our strategy did not outperform transformers, we believe the results are very promising, since it requires much less computational power than transformers. Moreover, our method requires less annotated data (only 10%) than transformers to achieve interesting results.

Our approach is available at `https://github.com/rafaelanchieta/toxic`.

6 Conclusion and Future Work

In this paper, we explored a semi-supervised strategy to deal with toxic comments from Twitter. We modeled the texts as a heterogeneous network graph with two node types and weighted edges among nodes. Then, we applied a regularization algorithm to extract features related to the toxic texts. Finally, we used these features to feed a classifier to identify and predict toxic comments. Our approach outperformed several graph-based methods and achieved a competitive result compared to the BERT model, using only 10% of the corpus. We hope that this graph model brings insights to hate speech detection research, helping to improve the results. Furthermore, our strategy may be employed in other languages, as it only requires an embedding representation.

As future work, we intend to explore the graph structure, analyzing some network measures, such as degree, centrality, community identification, and others. Also, we aim to examine contextual embeddings rather than traditional embeddings.

Acknowledgments

The authors are grateful to CAPES for supporting this work.

References

Steven Abney. 2007. *Semisupervised Learning for Computational Linguistics*, 1st edition. Chapman & Hall/CRC.

Rafael T Anchiêta, Rogério F de Sousa, and Thiago AS Pardo. 2020. Modeling the paraphrase detection task over a heterogeneous graph network with data augmentation. *Information*, 11(9):422.

Flor Miriam Plaza-del Arco, M Dolores Molina-González, L Alfonso Ureña-López, and M Teresa Martín-Valdivia. 2021. Comparing pre-trained language models for spanish hate speech detection. *Expert Systems with Applications*, 166:114120.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760, Perth, Australia. International World Wide Web Conferences Steering Committee.

Shubhanker Banerjee, Bharathi Raja Chakravarthi, and John P. McCrae. 2020. Comparison of Pretrained Embeddings to Identify Hate Speech in Indian Code-Mixed Text. In *Proceedings of the 2nd International Conference on Advances in Computing, Communication Control and Networking*, pages 21–25, Greater Noida, India. IEEE.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. 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*, pages 71–80, Amsterdam, Netherlands. IEEE.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Alexis CONNEAU and Guillaume Lample. 2019. Cross-lingual Language Model Pretraining. In *Advances in Neural Information Processing Systems*, page 11, Vancouver, Canada. Curran Associates, Inc.

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)*, pages 112–120, Osaka, Japan. The COLING 2016 Organizing Committee.

Vinicius Matheus Coutinho and Yuri Malheiros. 2020. Detecção de Mensagens Homofóbicas em Português no Twitter usando Análise de Sentimentos. In *Anais do IX Brazilian Workshop on Social Network Analysis and Mining*, pages 1–12, Porto Alegre, RS, Brasil. SBC.

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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Rehab Duwairi, Amena Hayajneh, and Muhannad Quwaider. 2021. A Deep Learning Framework for Automatic Detection of Hate Speech Embedded in Arabic Tweets. *Arabian Journal for Science and Engineering*, 46(4):4001–4014.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

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

Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Jie Gui, Rongxiang Hu, Zhongqiu Zhao, and Wei Jia. 2014. Semi-supervised learning with local and global consistency. *International Journal of Computer Mathematics*, 91(11):2389–2402.

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*, pages 122–131, Uberlândia, Brazil. Sociedade Brasileira de Computação.

Ming Ji, Yizhou Sun, Marina Danilevsky, Jiawei Han, and Jing Gao. 2010. Graph regularized transductive classification on heterogeneous information networks. In *In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 570–586, Barcelona, Spain. Springer.

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*, 2:1–14.

João Augusto Leite, Diego Silva, Kalina Bontcheva, and Carolina Scarton. 2020. Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924, Suzhou, China. Association for Computational Linguistics.

Peter Dias Paiva, Vanecy Matias da Silva, and Raimundo Santos Moura. 2019. Detecção automática de discurso de ódio em comentários online. In *Anais da VII Escola Regional de Computação Aplicada à Saúde*, pages 157–162, Teresina, Piauí, Brazil. Sociedade Brasileira de Computação.

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

Rogers 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*, pages 510–519, São Paulo, Brazil. Sociedade Brasileira de Computação.

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*, 55:477–523.

Hammad Rizwan, Muhammad Haroon Shakeel, and Asim Karim. 2020. Hate-speech and offensive language detection in Roman Urdu. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2512–2522, Online. Association for Computational Linguistics.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Rogério Figueredo de Sousa, Rafael Torres Anchiêta, and Maria das Graças Volpe Nunes. 2020. A graph-based method for predicting the helpfulness of product opinions. *iSys-Brazilian Journal of Information Systems*, 13(4):06–21.

Amarnag Subramanya and Partha Pratim Talukdar. 2014. Graph-based semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 8(4):1–125.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Tianchi Yang, Linmei Hu, Chuan Shi, Houye Ji, Xiaoli Li, and Liqiang Nie. 2021. Hgat: Heterogeneous graph attention networks for semi-supervised short text classification. *ACM Transactions on Information Systems*, 39(3).

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, pages 7370–7377, Honolulu, HI, USA. AAAI Press.

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*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

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*, page 793–803, New York, NY, USA. Association for Computing Machinery.

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*, pages 321–328, MA, USA.