Procesamiento del Lenguaje Natural, Revista nº 74, marzo de 2025, pp. 391-394 recibido 03-12-2024 revisado 27-01-2025 aceptado 01-02-2025

On the Keyword Extraction and Bias Analysis, Graph-based Exploration and Data Augmentation for Abusive Language Detection in Low-Resource Settings

Detección del lenguaje abusivo en entornos con escasos recursos mediante la extracción de palabras clave, el análisis de sesgos, la exploración basada en grafos y el aumento de datos

Gretel Liz De la Peña Sarracén
Universitat Politècnica de València
gredela@posgrado.upv.es

Abstract: PhD thesis in Computer Science focused on Natural Language Processing, written by Gretel Liz De la Peña Sarracén under the supervision of Prof. Paolo Rosso in the Universitat Politècnica de València (Spain). This thesis addresses the abusive language detection specifically emphasizing low-resource settings. The results highlight the potential of graph neural networks models and data augmentation techniques for improving this task. The thesis defense was held virtually on March 6th, 2024. The doctoral committee was composed by Dr. Arkaitz Zubiaga (Queen Mary University of London), Dra. Sara Tonelli (Fondazione Bruno Kessler), and Dra. Aitziber Atutxa (University of the Basque Country). An international mention was achieved, and the work was graded as excellent and awarded Cum Laude.
Keywords: Abusive Language Detection, Low-Resource Settings, Keyword Extraction, Bias Analysis, Graph-based Exploration, Data Augmentation.

Resumen: Tesis doctoral en Informática con enfoque en el Procesamiento del Lenguaje Natural realizada por Gretel Liz De la Peña Sarracén y dirigida por el Prof. Paolo Rosso en la Universitat Politècnica de València (España). Esta tesis aborda el análisis de la detección de lenguaje abusivo con un enfoque particular en entornos caracterizados por la escasez de datos. Los resultados destacan el potencial de los modelos basados en redes neuronales de grafos y técnicas de aumento de datos para el mejoramiento de esta tarea. La defensa de la tesis fue realizada de manera virtual el 6 de marzo de 2024 ante un tribunal compuesto por el Dr. Arkaitz Zubiaga (Queen Mary University of London), la Dra. Sara Tonelli (Fondazione Bruno Kessler), y la Dra. Aitziber Atutxa (University of the Basque Country). Se obtuvo la mención internacional y una calificación de sobresaliente Cum Laude.
Palabras clave: Detección del Lenguaje Abusivo, Extracción de palabras clave, Análisis de sesgos, Análisis basada en grafos, Aumento de datos.

1 Introduction

Abusive language detection is a task that has become increasingly important in the modern digital age, where communication occurs across numerous online platforms. The rise in online interactions has correspondingly led to a surge in instances of abusive language. Tackling this issue is essential for fostering a safe and inclusive online community. However, this endeavor encounters various challenges that render it a complex and continually evolving field of research and development. In particular, detecting abusive language in environments with sparse data poses an additional challenge, as the creation of accurate automated systems typically requires extensive annotated datasets.

Recent studies in abusive language detection focus on the problem of bias. These works reveal the presence of bias and its pos-

ISSN 1135-5948 DOI 10.26342/2025-74-27

sible impact on the task. We believe that this issue can increase when data is scarce, due to the limited availability of diverse and representative labeled data to train robust models. Beside, we found evidence that graph-based models possess considerable potential in low-resource settings. Such models enable the integration of minimal labeled data alongside a substantial volume of unlabeled data. Then, they provide several benefits for optimizing the use of scarce data by propagating information across a graph to make predictions.

Another observation in low-resource settings is that cross-lingual techniques can be highly effective. These techniques allow for the transfer of knowledge and models from rich-resource languages to low-resource languages, where models can make predictions in low-resource languages with minimal training examples. However, the linguistic diversity in low-resource languages may struggle to accommodate all language variations. Then, data augmentation is a valuable technique that can be designed to create new data that are specifically tailored to low-resource languages.

In this thesis, we investigated different aspects of abusive language detection, paying particular attention to environments with limited data. We conducted the study from three different aspects: 1) **analysis of bias toward abusive keywords in models**, 2) **graph-based models**, and 3) **data augmentation**. Considering these aspects, the research questions we aimed to answer in this thesis were:

RQ1 Could bias toward potential abusive keywords in the models affect the performance of abusive language detection in low-resource settings?

RQ2 What is the contribution of models based on graph neural networks for abusive language detection in low-resource settings?

RQ3 What is the contribution of data augmentation for abusive language detection in low-resource settings?

First, we studied the bias toward abusive keywords in models trained for abusive language detection. To this end, we proposed two methods for extracting potentially abusive keywords from datasets. We then evaluated the bias toward the extracted keywords and how this bias can be modified in order to influence abusive language detection performance. Second, we explored the application of graph neural network models for the detection of abusive language. On the one hand, we proposed a text representation framework designed to obtain a representation space that facilitates the clear differentiation of abusive texts from non-abusive ones. On the other hand, we assessed the effectiveness of convolutional graph neural network models to classify abusive texts. The next part of our research focused on analyzing how data augmentation can influence the performance of abusive language detection. To achieve this, we investigated two well-known techniques based on the principle of vicinal risk minimization and proposed a variant for one of these techniques.

2 Thesis Overview

The work presented in the thesis was organized in 8 chapters.

Chapter 1 This chapter is the introductory section, where we introduced the problems of abusive language, describing the issue of data limitation to detect it automatically.

The next 5 chapters were grouped in 3 main parts:

Part I: Keyword Extraction and Bias Analysis In this part, we presented the research concerning the bias associated with potential abusive keywords within the models and its implications for the efficacy of abusive language detection (**RQ1**).

Chapter 2 In this chapter, we proposed a method for extracting potentially abusive keywords, that leverages the ability of the multi-head self-attention mechanism of BERT to capture contextual and semantic information in texts. This research (De la Peña Sarracén and Rosso, 2021) was published in the journal Personal and Ubiquitous Computing.

Chapter 3 In this chapter, we presented a more straightforward approach for detecting potentially abusive keywords. This is a statistics-based method to identify prevalent keywords in hateful texts that are less common in other texts. We used this method to evaluate the bias toward abusive keywords and how it may affect the performance of abusive language detection. This method (De la Peña Sarracén and Rosso, 2023) was published in the journal Information Processing & Management.

Part II: Graph-Based Analysis In this part, we evaluated the potential of graph neural networks models in hate speech detection.

Chapter 4 In this chapter, we proposed a graph auto-encoder framework to obtain a latent representation from an initial text representation. We used this framework for hate speech detection by using the embeddings as input of a classifier. This work (De la Peña Sarracén and Rosso, 2022b) was published in the proceedings of the 13th Language Resources and Evaluation Conference.

Chapter 5 In this chapter, we studied a model based on convolutional graph neural networks to address mainly hate speech detection in scenarios with little data (**RQ2**). Our findings (De la Peña Sarracén and Rosso, 2022a) were published in the 27th International Conference on Applications of Natural Language to Information Systems.

Part III: Data Augmentation This part was comprised of **Chapter 6**, in which we addressed the issue of few-shot cross-lingual transfer learning in abusive language detection. We explored data augmentation techniques to deal with the issue of data scarcity that can lead to a high estimation error in few-shot learning (**RQ3**). These techniques are based on the principle of vicinal risk minimization that aims to increase the data in the vicinity of the few-shot samples. We explored two existing techniques: 1) SSMBA, which is based on a pair of functions to corrupt and reconstruct texts, and 2) MIXUP, which generates new samples from a linear combination of original instances pairs. Then, we proposed MIXAG, a variant of MIXUP, to parameterize the combination of instances with the angle between them. This work (De la Peña Sarracén et al., 2023) was published the proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Chapter 7 In this chapter, we complemented our study with further experiments to gain additional insights. First, we conducted new experiments to compare our two methods for keyword extraction; as well as extending the analysis of the bias toward the keywords extracted with both methods for particular low-resource settings. Then, we conducted an ablation analysis of different types of graphs neural networks for abusive language detection. Finally, we compared simple techniques with those based on vicinal risk minimization. Additionally, at the end of this chapter we included supplementary analyses for the study of potential spreaders of hate speech. We also presented our work published in the Conference and Labs of the Evaluation Forum (CLEF), where we contributed to the development of a dataset for profiling haters in Twitter, both in English and Spanish (Rangel et al., 2021). This is an overview of the shared task we assisted in organizing at PAN 2021.

Chapter 8 In this chapter, we outlined our contributions and provide insights into the open research lines for possible future works.

3 Conclusions and Contributions

Abusive language detection in an environment where data is scarce is a major challenge. The models often lack sufficient examples to grasp the subtleties of the language. Additionally, adapting models from resource-rich languages to those with limited resources is not trivial and can lead to diminished accuracy. The work presented in our thesis focused on the study of abusive language detection in low-resource settings considering three aspects.

First, we investigated how models can reflect existing biases in the training data that can lead to unfair detection of abusive language. We first found a set of terms where abusive detection might be biased, **proposing two methods to extract potentially abusive keywords from datasets**. One of the methods is based on the BERT attention mechanism and the other on statistics computed from word frequencies related to the class of abusive texts. Although the keywords extracted by the two methods did not overlap much, we found that both methods mainly extract abusive words. Then, we investigated the bias of the models toward these keywords. The experimental result of our research showed that modifying the bias toward potential abusive keywords in the models generally varies the performance of abusive language detection. We observed that the bias can be reduced when fine-tuning the models with abusive texts in which the keywords are not present and that this reduction can mean a performance improvement. However, in low-resource settings, it was not possible to determine how the bias needs to be varied in order to improve the performance.

The reason may lie in the lack of texts, which prevented a good fit of the model to appropriately mitigate the bias.

Second, we **assessed the role of graph neural networks models for abusive language detection**. According to our findings, **this type of models are promising for abusive language detection, especially in low-resource settings**. We found that they seem to have the potential to outperform transformer-based models for detecting abusive language. We also evaluated the suitability of graph auto-encoders to obtain a discriminatory representation between abusive and non-abusive texts.

Third, we **evaluated how the cross-lingual few-shot learning task can be enhanced with three data augmentation techniques**. We considered a model trained to detect abuse in English and fine-tuned it with some examples for another target language. Then, the number of examples in the target language was augmented with vicinal risk minimization techniques. An improvement in results was observed after using a model fine-tuned with this number of new samples. Thus, our results revealed that these techniques can be an effective strategy to improve abusive language detection in low-resource settings.

In conclusion, we asserted that the findings related to the research questions indicate that both **the use of graph neural networks models as well as data augmentation can lead to the improvement of abusive language detection in low-resource settings**. Besides, we noted that bias mitigation can fail when the data is scarce, highlighting the need to explore alternative approaches in such contexts. Finally, we introduced a study on abusive language detection from the user's perspective. We found that it is possible to automatically identify potential haters based on a stream of publications. Furthermore, we noted that important insights can be gained from analyzing the graph of relationships between users. In particular, we observed that a few users who play a crucial role in connecting different parts of the network may be the most important to identify to prevent the spread of abusive messages.

Acknowledgments

This work was supported by various financial projects. Among them: Fairness and Transparency for equitable NLP applications in social media, funded by MCIN/AEI/10.13039/501100011033 and ERDF, EU A way of making EuropePI; and MISMIS-FAKEnHATE on Misinformation and Miscommunication in social media: FAKE news and HATE speech (PGC2018-096212-B-C31).

References

De la Peña Sarracén, G. L. and P. Rosso. 2021. Offensive Keyword Extraction based on the Attention Mechanism of BERT and the Eigenvector Centrality using a Graph Representation. *Personal and Ubiquitous Computing*, pages 1–13.

De la Peña Sarracén, G. L. and P. Rosso. 2022a. Convolutional Graph Neural Networks for Hate Speech Detection in Data-Poor Settings. In *27th International Conference on Applications of Natural Language to Information Systems, NLDB 2022, Valencia, Spain, June 15–17, 2022, Proceedings*, pages 16–24.

De la Peña Sarracén, G. L. and P. Rosso. 2022b. Unsupervised Embeddings with Graph Auto-Encoders for Multi-domain and Multilingual Hate Speech Detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2196–2204.

De la Peña Sarracén, G. L. and P. Rosso. 2023. Systematic Keyword and Bias Analyses in Hate Speech Detection. *Information Processing & Management*, 60(5):103433.

De la Peña Sarracén, G. L., P. Rosso, R. Litschko, G. Glavaš, and S. P. Ponzetto. 2023. Vicinal risk minimization for few-shot cross-lingual transfer in abusive language detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4069–4085.

Rangel, F., G. L. De la Peña Sarracén, M. A. Chulvi-Ferriols, E. Fersini, and P. Rosso. 2021. Profiling Hate Speech Spreaders on Twitter Task at PAN 2021. In *Proceedings of the Working Notes of CLEF 2021, Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st to 24th, 2021*, pages 1772–1789. CEUR.