Hate Speech Detection Using NLP
An Annotated Bibliography

Danel Levi Bacarra (dlbacarra@up.edu.ph) Marfred James Deen (mpdeen@up.edu.ph) Arwin Delasan (avdelasan@up.edu.ph) Jed Edison Donaire (jjdonaire@up.edu.ph) Princess Mae Parages (pbparages@up.edu.ph) Sheldon Arthur Sagrado (smsagrado@up.edu.ph) James Torres (jtorres1@up.edu.ph)

3 October 2024

References

[1] A. Al-Hassan and H. Al-Dossari, “Detection of hate speech in social networks: A survey on multilingual corpus,” 02 2019, pp. 83–100.

> This paper by Al-Hassan and Al-Dossari provides a comprehensive survey of hate speech detection within social networks, with a specific focus on multilingual corpora. It explores the challenges posed by language diversity and the need for robust datasets capable of handling hate speech in multiple languages. The paper reviews various techniques and tools used for detecting hate speech in different languages, emphasizing the importance of addressing cultural and contextual nuances. It also highlights gaps in existing research, including limitations in available datasets and the need for more refined tools to handle multilingual contexts effectively.
>
> In relation to the paper by Jahan and Oussalah, this survey is highly relevant as it complements the systematic review of hate speech detection using natural language processing (NLP) and deep learning. Al-Hassan and Al-Dossari’s focus on multilingual corpora highlights a crucial area of ongoing research, especially given the limitations noted by Jahan and Oussalah in current hate speech detection models. The multilingual aspect

covered in this reference enriches the understanding of how NLP models might be adapted or improved to work across diverse linguistic landscapes, contributing to the future research directions outlined in the 2023 paper.

[2] S. S. Aluru, B. Mathew, P. Saha, and A. Mukherjee, “Deep learning models for multilingual hate speech detection,” *CoRR*, vol. abs/2004.06465, 2020. [Online]. Available: https://arxiv.org/abs/2004.06465

The paper addresses the challenge of multilingual hate speech detection by analyzing datasets from 9 languages across 16 sources to determine the most effective models for varying levels of data availability. In low-resource settings with limited training data, simpler models like LASER embedding with logistic regression perform best, while BERT-based models excel when more data is available. The study also explores zero-shot classification, finding that languages like Italian and Portuguese achieve good results even without direct training data. The paper’s key contribution is a ”recipe catalog” that offers guidance on selecting the most suitable model based on data availability, demonstrating that LASER + LR is optimal for low-resource scenarios , while BERT-based models are more effective with abundant data.
This research is important for advancing automatic hate speech detection by analyzing multilingual datasets, extending beyond the usual English focus to include 9 languages. It compares models like LASER with logistic regression and BERT-based models, identifying the most effective methods for both low-resource and high-resource scenarios. The ”recipe catalog” provides practical guidance for choosing the right model based on data availability, aiding researchers in handling languages with less data. By sharing their code and setup, the authors support future research, promoting more inclusive and effective hate speech detection systems globally.

[3] M. Di Capua, E. Di Nardo, and A. Petrosino, “Unsupervised cyber bullying detection in social networks,” in *2016 23rd International Conference on Pattern Recognition (ICPR)*, 2016, pp. 432–437.

In this paper, the authors address the growing issue of cyberbullying among young people who have grown up in a digital

world. They define cyberbullying as ”willful and repeated harm inflicted through the use of electronic devices.” To tackle this problem, they propose an automatic detection method using techniques from Natural Language Processing (NLP) and machine learning. Their model is based on Growing Hierarchical Self-Organizing Maps (GHSOMs), which helps to cluster text documents that contain signs of bullying by analyzing both the meaning and structure of the language used.
The authors tested their model on Twitter, YouTube, and Formspring, demonstrating that their unsupervised approach can effectively identify instances of cyberbullying with good performance metrics. The results of their study indicate that the proposed method achieves an accuracy rate of approximately 86in detecting cyberbullying across different platforms. This is significant given the challenges of collecting labeled data, as traditional supervised models often struggle due to the sporadic nature of bullying messages.
The authors emphasize that their unsupervised model not only offers promising results but also has the potential for real-world applications in monitoring social media for bullying behavior. By utilizing features like syntactic, semantic, sentiment , and social characteristics, their approach highlights the importance of considering multiple aspects of communication in the detection process. Overall, this paper makes a valuable contribution to the understanding of automated cyberbullying detection and suggests a pathway for future research and application in this important area.

[4] R. Gomez, J. Gibert, L. Gómez, and D. Karatzas, “Exploring hate speech detection in multimodal publications,” *CoRR*, vol. abs/1910.03814, 2019. [Online]. Available: http://arxiv.org/abs/1910.03814

In ”Exploring Hate Speech Detection in Multimodal Publications,” Gomez, Gibert, Gomez, and Karatzas (2020) tackle the complex issue of detecting hate speech in multimodal content, focusing on Twitter posts that combine text and images. They introduce and annotate a large-scale dataset, MMHS150K, which includes 150,000 tweets with multimodal components, to explore how combining textual and visual information affects hate speech detection. The authors evaluate several models that jointly analyze both text and images and compare them to models that rely solely on text.

Their findings show that while images can provide additional context for detecting hate speech, multimodal models do not outperform unimodal text-based models. Despite this, the authors provide valuable insights into the potential of multimodal detection, noting the challenges posed by the complexity and diversity of the relations between visual and textual data. The dataset, which they make publicly available, and their analysis of various detection methods , contribute significantly to advancing research in hate speech detection, especially in the context of social media. This paper is relevant to researchers in fields such as machine learning, natural language processing, and social media analysis, as it opens new avenues for studying the interaction between visual and textual data in online content moderation.

[5] A. Jha and R. Mamidi, “When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data,” in *Proceedings of the Second Workshop on NLP and Computational Social Science*, D. Hovy, S. Volkova, D. Bamman, D. Jurgens, B. O’Connor, O. Tsur, and A. S. Doğruöz, Eds. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 7–16. [Online]. Available: https://aclanthology.org/W17-2902

The paper investigates the phenomenon of ambivalent sexism using Twitter data, focusing on hostile and benevolent sexist inclinations. Hostility in sexism entails clearly negative attitudes, conversely benevolent sexism, though appearing favorable , reinforces traditional gender norms. The researchers assembled a collection of tweets to label them as ”Hostile,” ”Benevolent,” or ” Others” and used machine learning methods like Support Vector Machines (SVM), Sequence-to-Sequence models for analysis. This analysis focuses on the subtlety of benevolent sexism and its influence on social equity despite its promising appearance, benevolent sexism is harmful to women’s mental capacity.

This research is important for understanding the nuances of sexism on social media. This paper highlights ambivalent sexism to address a gap in existing research that mainly focused on hostile sexism. The authors applied a thorough methodology consisting of creating a specialized dataset and using multiple machine learning techniques for classification. A drawback of the analysis is the manual annotation method that might influence the results. The paper's strength lies in its comprehensive dataset and comparison of machine learning models, providing insights for further research in computational social science.

[6] S. Malmasi and M. Zampieri, "Challenges in discriminating profanity from hate speech," *CoRR*, vol. abs/1803.05495, 2018. [Online]. Available: http://arxiv.org/abs/1803.05495

The study tackles the issue of distinguishing common profanity from hate speech on social media with a supervised classification approach. A novel dataset particularly annotated for this task was utilized, incorporating features such as n-grams, skip-grams, and clustering-based word representations. The study obtained 80three-class classification task, indicating that current methodologies encounter difficulties with this issue due to the subjective nature of annotations and the overlap between profanity and hate speech. The research emphasizes the necessity for more profound language traits to encapsulate context, while superficial features are inadequate. The research indicates that enhancing annotation quality and employing ensemble classifiers might improve performance in further studies.
The study enhances hate speech detection by illustrating the intricacies involved in differentiating between common profanity and hate speech, particularly inside social media environments. It emphasizes the shortcomings of conventional methods that depend on superficial characteristics such as n-grams by employing sophisticated classifiers and ensemble techniques. The research indicates that hate speech detection algorithms must consider the subtle distinctions between offensive words used for emphasis and targeted attacks, which is essential for minimizing false positives. It underscores the necessity of comprehensive language and contextual investigation to improve the precision of automated hate speech identification. The study's findings on classifier performance and

annotator bias provide guidance for enhancing hate speech detection in the future.

[7] H. Watanabe, M. Bouazizi, and T. Ohtsuki, “Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection,” *IEEE Access*, vol. 6, pp. 13 825–13 835, 2018.

> This article tries to discern hate speech on Twitter by employing a novel method. This method tries to consider hate speech’s sentiment, semantics, unigrams and speech patterns. The authors’ experiments show that combining these macro features improves binary classification of hateful/offensive tweets against clean/non offensive tweets. However, performance dwindled once a deeper discernation between hateful and offensive tweets (i.e. ternary classification).
> The value in Watanabe et al.’s paper lies in introducing a pragmatic framework for collecting a set of expressions in online social networks. There is potential in applying this method for other kinds of expression. However, the authors did point out that there might be a need in building a rich dictionary of patterns related to that expression, which in their case was hate.