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.

> The paper by Al-Hassan and Al-Dossari provides a thorough survey of hate speech detection on social networks, with a special focus on multilingual datasets. It discusses the challenges that come with language diversity and highlights the need for strong datasets that can effectively identify hate speech in different languages. Furthermore, the authors review various techniques and tools used for detecting hate speech, stressing the importance of considering cultural and contextual factors. Additionally, they point out gaps in current research, such as the limitations of existing datasets and the need for better tools to handle multilingual situations effectively.
>
> Moreover, this survey is particularly important because it adds to previous research, complementing systematic reviews of hate speech detection using natural language processing (NLP) and deep learning. By emphasizing multilingual datasets, Al-Hassan and Al-Dossari address a crucial area of research, especially given the limitations found in current hate speech detection models. Their focus on multilingual issues improves

our understanding of how NLP models can be adapted or enhanced to work across different languages, thus contributing to future research.

[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

This paper addresses the problem of detecting hate speech in multiple languages. Specifically, it examines data from 9 languages collected from 16 sources to find the best models for different amounts of training data. In situations with limited training data, simpler models like LASER embedding combined with logistic regression work well. However, when there is more data, BERT-based models perform better.
Moreover, the study also looks into zero-shot classification, which allows some languages, such as Italian and Portuguese, to achieve decent results even without specific training data. A major contribution of this paper is a ” recipe catalog” that helps researchers choose the best model depending on the available data. Notably, it shows that using LASER with logistic regression is ideal for low-resource scenarios, while BERT-based models are more effective when there is plenty of data. Furthermore, this research is significant because it improves automatic hate speech detection by focusing on multilingual datasets, not just English. It compares various models, highlighting the best approaches for both low-resource and high-resource situations. Additionally, the ”recipe catalog” offers practical advice for selecting the right model based on data availability, which can help researchers work with languages that have less data. Finally, the authors share their code and setup to support future research, encouraging the development of more inclusive and effective hate speech detection systems worldwide.

[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. Specifically, 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.
Moreover, 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. In fact, the results of their study indicate that the proposed method achieves an accuracy rate of approximately 86showcasing its effectiveness in detecting cyberbullying across different platforms. This finding is significant, given the challenges of collecting labeled data, as traditional supervised models often struggle due to the sporadic nature of bullying messages.
Furthermore, 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 system-based 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 et al. (2020) examine the detection of hate speech using a multimodal approach, combining both textual and visual information. They focus on the significance of integrating various forms of content to achieve more accurate hate speech detection in social media posts that contain images alongside text. The authors argue that traditional text-based approaches may not adequately capture the full context of hate speech, given the prevalent use of images and other media in online communication.

The research utilizes a dataset containing multimodal posts from Twitter, allowing the authors to assess the effectiveness of their proposed methods. The study employs advanced machine learning techniques, including convolutional neural networks (CNNs), to analyze the visual components of posts, along with natural language processing for the textual data. As a result, the study demonstrates improved detection accuracy when both text and images are considered, illustrating the need for a more holistic approach to hate speech identification.

Moreover, the findings highlight the potential for enhancing hate speech detection systems by incorporating multiple modalities, offering insights for future research and application. By examining the interplay between textual and visual elements in online content, this study contributes to a more comprehensive understanding of hate speech in the digital age, underlining the importance of interdisciplinary methodologies for addressing complex social issues.

[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 ambivalent sexism using Twitter data, specifically looking at two types: hostile and benevolent sexism. Hostile sexism involves openly negative attitudes, while benevolent sexism, despite seeming positive, actually reinforces traditional gender norms. The researchers collected tweets and categorized them as ”Hostile,” ”Benevolent,” or ”Others.” Furthermore, they employed machine learning techniques, such as Support Vector Machines (SVM) and Sequence-to-Sequence models, for their analysis. This study highlights the complexities of benevolent sexism and its impact

on social equity, showing that it can be harmful to women's mental health, even though it may appear positive at first.
Moreover, this research is crucial for understanding the details of sexism on social media. It focuses on ambivalent sexism to fill a gap in existing studies, which have largely concentrated on hostile sexism. The authors used a detailed methodology, creating a specialized dataset and applying various machine learning techniques for classification. However, a limitation of the study is the manual annotation process, which could affect the results. Overall, the paper's strengths include its comprehensive dataset and the comparison of different machine learning models, offering valuable insights for future 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 challenge of distinguishing common profanity from hate speech on social media using a supervised classification approach. To achieve this, a unique dataset was utilized, incorporating features such as n-grams, skip-grams, and clustering-based word representations. Notably, the study achieved 80classification task. This indicates that current methodologies face difficulties due to the subjective nature of annotations and the overlap between profanity and hate speech. Furthermore, the research emphasizes the need for deeper linguistic traits to capture context, as superficial features are inadequate. It suggests that improving annotation quality and using ensemble classifiers might enhance performance in future studies.
In addition, the study enhances hate speech detection by illustrating the complexities involved in differentiating between common profanity and hate speech, particularly in social media environments. It highlights the shortcomings of conventional methods that rely on superficial characteristics, such as n-grams, and instead employs advanced classifiers and combined techniques. Moreover, the research shows 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. Overall, it emphasizes the importance of comprehensive language and contextual analysis to improve the accuracy of automated hate speech identification. Finally, the study's findings on classifier performance and annotator bias provide valuable guidance for advancing 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 investigates hate speech on Twitter using an innovative method that incorporates various dimensions of hate speech, including sentiment, semantics, unigrams, and speech patterns. The authors conducted extensive experiments, revealing that the integration of these macro features significantly enhances the binary classification of hateful or offensive tweets compared to clean or non-offensive tweets. However, the study found that performance declined when attempting to differentiate between hateful and offensive tweets in a ternary classification framework, indicating the complexities involved in accurately categorizing such expressions.

In response to these challenges, the authors propose a pragmatic framework for collecting and analyzing a diverse range of expressions within online social networks. This framework effectively addresses the challenge of hate speech detection and lays the groundwork for applying similar methodologies to identify other forms of harmful communication, such as bullying or misinformation. Furthermore, a key recommendation from the study is the development of a comprehensive dictionary of linguistic patterns related to hate speech, which the authors argue is essential for improving the effectiveness of their detection approach.

Overall, Watanabe et al.'s research offers valuable insights into online safety and lays the groundwork for future studies aimed at reducing the impact of hate speech on social media platforms. Ultimately, their findings enhance our understanding of hate speech dynamics on Twitter and provide practical recommendations for improving user safety and promoting healthier online interactions.