The Investigation of Progress Related to Harmful Speech Detection Models

Ruikun Wang[a]
Software Engineering, Xiamen University Malaysia, Sepang, Malaysia

Keywords: Offensive Language Detect, Machine Learning, Artificial Intelligence.

Abstract: With the rise of the internet, individuals can express their opinions or engage in conversations with others on social platforms. However, along with this development comes the proliferation of harmful speech on these platforms. Harmful speech poses various dangers, such as fueling conflicts and contributing to social issues. Consequently, effective detection and regulation of harmful speech have become hot topics of discussion. This paper provides an overview of several existing models for detecting harmful speech. Firstly, it reviews traditional detection methods and introduces three classic detection models: the Dictionary method, which identifies harmful vocabulary; the n-gram method and skip-gram method, which assess context for detection. Additionally, it reviews machine learning methods, highlighting both traditional machine learning approaches and the recent popular use of large language models as examples. Through analysis, it is noted that traditional detection methods exhibit low implementation costs but suffer from questionable accuracy due to challenges in understanding natural language, leading to reduced precision. In contrast, traditional machine learning methods, although capable of to some extent understanding natural language, require significant human and material resources for model training. Meanwhile, models utilizing large language datasets are able to further comprehend natural language, enhancing model accuracy. The article points out potential shortcomings in current models, such as inaccuracies in detection results due to diverse cultural backgrounds, misidentification of emerging words and changes in word meanings, as well as negative impacts of data biases on model performance.

1 INTRODUCTION

As Internet technology advances, social platforms are increasingly integrated into people's daily lives. While they serve to bring individuals closer together, these platforms also foster the dissemination and escalation of hate speech. Hate speech, as defined by UNESCO, is characterized by the expression of discriminatory or disparaging remarks directed towards specific person or groups originating on attributes for instance gender, faith, ethnicity, or race (Iginio et al., 2015; Lee et al., 2018). Such speech targets specific social or demographic groups and encourages harm or incites violence towards them. The dissemination of these harmful messages incites emotional responses and fosters conflicts among individuals, potentially resulting in severe repercussions. Therefore, efficiently and accurately identifying and addressing harmful speech on social platforms is of significant importance to prevent its negative impact.

Consequently, the central focus of current research initiatives lies in efficiently and accurately identifying and addressing harmful speech on social platforms to prevent its negative impacts. In the early stages, scholars attempted to identify harmful speech using dictionary-based methods (Lee et al., 2018) and rule-based methods (Chen et al., 2012). An example of this approach is the classification of a sentence as hate speech based on the presence of a second-person pronoun and derogatory term. However, this method inherently requires the expertise of domain specialists and entails substantial manpower and resource allocation, thus limiting its generalizability.

What's more, this method does not enable machines to recognize the essential characteristics of harmful speech. People must continuously update dictionaries with new inappropriate vocabulary.

[a] https://orcid.org/0009-0008-0630-2962

Wang, R.
The Investigation of Progress Related to Harmful Speech Detection Models.
DOI: 10.5220/0012902100004508

In *Proceedings of the 1st International Conference on Engineering Management, Information Technology and Intelligence (EMITI 2024)*, pages 57-61
ISBN: 978-989-758-713-9

Scholars have explored another approach involving the utilization of deep learning in detecting harmful speech (Wang et al., 2020). By employing deep learning models and neural networks (Qiu et al., 2024), scholars are able to extract features of harmful speech, abstract the harmful nature of a sentence, and subsequently classify the text. In addition, deep learning models can leverage contextual cues to further distinguish sentence information. This model has the ability to extract high-dimensional semantic features of harmful speech from deeper layers, demonstrating a certain degree of transferability. Nonetheless, a significant limitation of this approach is the substantial time and manpower required for its implementation.

The aforementioned methods tend to concentrate heavily on extracting semantic features at the word level, often overlooking spelling features at the character level (Zhou et al., 2024). Users have been observed replacing pejorative words in sentences with homophones or similar sounding words, or deliberately misspelling words to circumvent detection of prohibited terms. In some cases, users may purposefully misspell specific words to give emphasis to certain points. Furthermore, implicit harmful speech, such as irony that subtly conveys emotions like disgust, could be misidentified, resulting in subpar model performance. Caselli et al. noted that pre-trained language models encounter challenges in accurately identifying implicit harmful speech (Caselli et al., 2020).

Some scholars have proposed using n-gram features to address the issue of model inaccuracies caused by spelling errors. N-gram, which is based on N words or characters in the text, explores the semantics of sentences and words. By doing so, it partially mitigates the problem of model inaccuracies due to spelling errors. This method offers a more detailed analysis of the text, enhancing the precision of models utilized in natural language processing.

In recent years, scholars have proposed large language models to apply deep learning in natural language processing for detecting harmful speech. Their primary objective is to comprehend and produce human language. By undergoing extensive training on vast quantities of text data, these large language models can acquire diverse patterns and structures of human language, allowing for a more efficient and precise detection of harmful speech.

2 METHOD

2.1 Traditional Detection

Researchers divide text detection into word-level detection and sentence-level detection as the most fundamental aspect of text detection. For word-level detection, researchers often utilize character n-grams and employ dictionary-based methods. Moreover, researchers frequently analyze the relationships between words by utilizing word n-grams and word skip-grams. More details about them are provided below.

2.1.1 Dictionary

One of the most fundamental methods for identifying hate speech is by analyzing the connotations of words to assess potential harm in a statement. This word-level detection technique serves as a cornerstone in hate speech detection efforts. To implement this approach, researchers commonly develop dictionaries containing harmful words. For instance, Razavi et al. compiled a dictionary of insults and derogatory terms to evaluate the level of aggressiveness in sentences (Razavi et al., 2010). Consequently, Njagi et al. further explored and validated the efficacy of this method in hate speech detection (Gitari et al., 2015).

2.1.2 N-Gram

In natural language processing, researchers often utilize n-grams, which are sequences composed of n adjacent symbols (letters or words) arranged in a specific order, to examine the context of sentences. The word-level n-gram divides words into multiple groups for mutual understanding at the word level. It is the most commonly referenced one for feature extraction from corpora. As highlighted by Davidson, T. et al., the use of word n-gram features has been shown to produce favorable outcomes when identifying hate speech (Davidson et al., 2017).

Character n-grams have gained significant attention from researchers for their effectiveness in detecting misspelled words on social media, which can either be unintentional or intentional and affect judgment. Köffer et al. successfully used character 2-grams and character 3-grams as features to identify hate speech, resulting in notable accuracy (Köffer et al., 2018). Additionally, Sandaruwan et al. highlighted the excellence of Character n-gram features in identifying spelling errors and issues related to similar character substitutions (Sandaruwan et al., 2019).

2.1.3 Word Skip-Gram

In analyzing word structures, word skip-gram and word n-gram exhibit similarities, yet differ in their respective approaches. For example, when considering the sentence "I bought a watch yesterday afternoon" a 1-skip gram analysis would yield the following features: "I a," "bought watch," "a yesterday," and "watch afternoon" (Sandaruwan et al., 2019). This technique is adept at extracting essential words while disregarding structural words. However, a notable limitation of 1-skip gram is its tendency to overlook non-structural components of the sentence.

2.2 Machine Learning-Based Detection

Machine learning is an algorithm that learns from data to perform tasks without explicit instructions and can generalize to unseen data. It enables algorithms to understand natural language and determine the potential harm of statements.

2.2.1 Traditional Machine Learning

In researching hate speech detection, researchers commonly turn to supervised learning algorithms in traditional machine learning. These algorithms function by analyzing datasets that have been annotated by humans using different models, allowing machines to develop effective hate speech detection models. The literature strongly establishes the credibility of using machine learning for hate speech detection (Khan & Qureshi, 2022; Ates et al., 2021; Khan et al., 2021). In their research, Emre Cihan et al. utilized models such as Naive Bayes, Gaussian Naïve Bayes, LDA, QDA, and LGBM to develop a detection model. However, a notable drawback of this approach is the requirement for a sufficiently large corpus to achieve optimal results, necessitating a considerable investment of human resources.

2.2.2 Large Language Model

The mention of large language models often brings ChatGPT to mind. Large language model is a kind of machine learning model. It is aiming for the generation and comprehension of nature language. These models have larger corpora and more parameters, allowing them to grasp natural language intricacies from higher dimensions and closely mirror human cognition. Compared to traditional machine learning approaches, large language models do not necessitate labeled data; they simply require an ample amount of data and parameters, thereby diminishing the reliance on human intervention. Recent research by Garani et al. demonstrates that integrating large language models into systems for harmful speech detection can greatly enhance accuracy (Yogita Garani et al., 2023). Their study leveraged models such as ChatGPT and XLM-Roberta to bolster the precision of harmful speech detection mechanisms.

3 RESULTS AND DISCUSSIONS

Traditional methods tend to be less complex but also less accurate than newer approaches, with a focus on word-level and character-level features that may hinder their performance in understanding natural language. The classic Dictionary method involves simply checking whether users have used prohibited words for detection. This method is the simplest and fastest, although it often suffers from issues such as the emergence of new prohibited words and misspelled prohibited words, leading to lower accuracy. Furthermore, this method solely relies on the prohibited word list for detection when encountering neutral terms, making it prone to false positives for some neutral words.

N-grams and word skip-grams are effective techniques that partially tackle the challenge of identifying misspelled prohibited words. These approaches demonstrate a degree of contextual understanding in sentences and vocabulary, which aids in the detection of misspelled prohibited words. Nevertheless, the limitation lies in their inability to completely grasp the intricacies of natural language.

Machine learning can better understand natural language, fundamentally distinguishing between harmful and normal discourse by comprehending natural language at the sentence level. Users have the ability to enhance the accuracy of detection models by selecting different models. However, a significant amount of human effort is required for the manual evaluation of all training data in machine learning. Moreover, models generated by traditional machine learning methods can only be used for the corresponding language, necessitating retraining for other languages. Researchers achieved promising results, with the accuracy of the LGBM model reaching over 90% (Ates et al., 2021).

Utilizing large language models enables individuals to construct more advanced hate speech detection systems based on existing large-scale models, for example, ChatGPT. LLM is able to have a deeper understanding of natural language and a higher-dimensional comprehension of its meaning allows for a fundamental distinction of whether the language is harmful. This distinction is key in developing effective hate speech detection systems for combatting harmful content online. Based on Yogita Garani et al.'s research, ChatGPT outperformed other LLMs with an accuracy rate of 95.25%, compared to approximately 70% accuracy for other models. This difference in performance could be attributed to biases in the dataset. (Yogita Garani et al., 2023).

Current hate speech detection systems have shown some limitations despite their advancements shown in Table 1. Large language models, for example, may struggle with understanding different cultural contexts, resulting in potential misjudgements. Moreover, the evolution of harmful speech on the internet poses a challenge as new forms emerge over time, and the connotations of certain words may alter from positive to negative. Consequently, addressing these drawbacks often involves the manual addition of new data as an interim solution, emphasizing the importance of researchers to explore effective strategies within this scope. Moreover, biases and imbalances in datasets can impact models, resulting in additional misjudgements about specific groups or topics. It is crucial to address methods for mitigating these data biases.

Table 1: The characteristics, advantages and disadvantages of the three method.

Method	Characteristics	Advantages	Dis-advantages
Classic Method	Involves using dictionary and N-grams techniques to check for prohibited words and to improve partial contextual under-standing.	Provides basic and quick identification of misspelled banned words, with some consideration of context.	Lower accuracy, vulnerability to new banned terms and misspellings, and susceptibility to false positives.
Machine Learning	Users can improve the accuracy of under-standing natural language by selecting different models that distinguish harmful from normal discourse at the sentence level.	Under-standing natural language at the sentence level enhances detection accuracy.	Significant human effort is needed for manual evaluation of training data and language-specific models.
Large Language Models	Improved under-standing of natural language, enhanced comprehension in higher dimensions, and clear recognition of harmful language.	Facilitates sophisticated hate speech detection systems and promotes a more in-depth comprehension	May struggle to understand diverse cultural contexts and may need to manually add new data.

4 CONCLUSION

This paper presents an overview of research methods aimed at detecting harmful speech on social platforms, classifying the existing methodologies into traditional methods, machine learning methods, and techniques based on LLM. Traditional methods, such as Dictionary, n-gram, and skip-gram approaches, are known for their simplicity and ease of implementation. However, they tend to exhibit lower accuracy levels due to issues like spelling errors and limited corpus sizes. On the contrary, machine learning techniques demonstrate a deeper comprehension of natural language, albeit necessitating substantial data labeling efforts and encountering hurdles in language transfer. In contrast, methods leveraging large language models offer the potential for heightened natural language understanding, thus enabling more precise and efficient identification of different types of speech.

Existing detection systems have limitations, as they may misjudge due to their inability to understand the cultural backgrounds of different communities. The evolution of harmful speech on the internet, along with the emergence of new forms and changes in the connotations of vocabulary over time, also poses challenges to existing detection systems. Dataset biases can impact the models. This paper highlights directions for research in the harmful speech detection systems field that need to be improved. It should be noted that this paper does not cover all research methods and notes that the accuracy of models may vary based on the selection of machine learning and large language models. Future plans involve incorporating additional models such as Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), among others.

REFERENCES

Ates, E. C., Bostanci, E., & Güzel, M. S.,2021. Comparative Performance of Machine Learning Algorithms in Cyber-bullying Detection: Using Turkish

Language Pre-processing Techniques. CoRR, vol. abs/2101.12718.

Caselli, T., Basile, V., Mitrovic, J., Kartoziya, I., & Granitzer, M., 2020. I Feel Offended, Don't Be Abusive!: Implicit/Explicit Messages in Offensive and Abusive Language. 6193–6202.

Chen, Y., Zhou, Y., Zhu, S., & Xu, H., 2012. Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing.

Davidson, T., Warmsley, D., Macy, M., & Weber, I., 2017. Automated Hate Speech Detection and the Problem of Offensive Language. Proceedings of the International AAAI Conference on Web and Social Media, 11(1), 512–515.

Gitari, N. D., Zhang, Z., Damien, H., & Long, J., 2015. A Lexicon-based Approach for Hate Speech Detection. International Journal of Multimedia and Ubiquitous Engineering, 10(4), 215–230.

Iginio Gagliardone, Gal, D., Alves, T., Martinez, G., & Unesco., 2015. Countering online hate speech. United Nations Educational, Scientific And Cultural Organization.

Khan, M. M., Shahzad, K., & Malik, M. K., 2021. Hate Speech Detection in Roman Urdu. ACM Transactions on Asian and Low-Resource Language Information Processing, 20(1), 1–19.

Khan, S., & Qureshi, A., 2022. Cyberbullying Detection in Urdu Language Using Machine Learning. https://doi.org/10.1109/etecte55893.2022.10007379

Köffer, S., Riehle, D. M., Steffen Höhenberger, & Becker, J., 2018. Discussing the Value of Automatic Hate Speech Detection in Online Debates.

Lee, H.-S., Lee, H.-R., Park, J.-U., & Han, Y.-S., 2018. An abusive text detection system based on enhanced abusive and non-abusive word lists. Decision Support Systems, 113, 22–31.

Qiu, Y., Hui, Y., Zhao, P., Cai, C. H., Dai, B., Dou, J., ... & Yu, J. 2024. A novel image expression-driven modeling strategy for coke quality prediction in the smart cokemaking process. Energy, 130866.

Sandaruwan, H. M. S. T., Lorensuhewa, S. A. S., & Kalyani, M. A. L., 2019, September 1. Sinhala Hate Speech Detection in Social Media using Text Mining and Machine learning. IEEE Xplore.

Wang, K., Lu, D., Soyeon Caren Han, Long, S., & Poon, J., 2020. Detect All Abuse! Toward Universal Abusive Language Detection Models. ArXiv.

Yogita Garani, Joshi, S., & Kulkarni, S., 2023. Offensive Sentiment Detection with Chat GPT and Other Transformers in Kannada.

Zhou, X., Fan, X., Yang, Y., Diao, Y., & Ren, G., 2024. Detection of Inappropriate Comments Based on Semantic Spelling Understanding and Gated Attention Mechanism (In Chinese). Computer Applications and Software, 01, 112-118+125.