Report By: Neha Gode & Hatim Sawai

Automated Cyberbullying Detection on Social Networks

Introduction

The rise of social media (like Facebook, Twitter) has made it easier for people to connect. But with this good comes a bad side: cyberbullying. This is bullying that happens online, through texting, social media, and other platforms. It can have serious effects, making people feel bad about themselves or even leading to suicide. To fight this, researchers are using powerful computer tools to automatically detect cyberbullying when it happens online.

What algorithms have been used?

Various algorithms have been employed in the development of cyberbullying detection systems. These algorithms typically follow a machine learning pipeline, starting with data collection from social media platforms. Data can be annotated through keyword filtering or manual labeling via crowd-sourcing platforms like Appen or Amazon Mechanical Turk. Inter-annotator agreement scores, measured using methods like Cohen's kappa or Krippendorff's alpha, help assess the reliability of annotations.

To address dataset imbalances, data sampling techniques are applied which helps to reduce biases and ensure that machine learning models learn to distinguish between cyberbullying and non-cyberbullying cases effectively.. Pre-processing steps include natural language processing tasks such as part-of-speech tagging**.** Removing stop words is a standard step but incase of cyberbullying second and third nouns could be an important indicator to categorize the text. Feature selection involves extracting user features (e.g., gender, age) and social media metadata (e.g., likes, comments) alongside textual, sentiment, and emotion features. Traditional machine learning models like TFIDF, coupled with sentiment and contextual features, have been utilized for cyberbullying detection, often yielding better results with binary classification models over multiclass approaches.

Deep learning models such as Convolutional Neural Networks (CNNs) have also been explored, particularly for detecting cyberbullying across multiple social platforms. Transfer learning approaches, although less common, have shown promise, as seen in the XP-CB framework, which fine-tunes the BERT Model for cross-platform configurations. Performance evaluation metrics include accuracy, F1-score, precision, recall, and Receiver Operating Characteristic-Area Under the Curve (ROC-AUC) scores, computed based on true positives, true negatives, false positives, and false negatives. Overall, these methodologies aim to effectively identify and mitigate instances of cyberbullying in online spaces.

Model optimization strategies, such as transfer learning approaches like the XP-CB framework, aim to enhance performance across different social platforms and linguistic variations. By fine-tuning pre-trained models like BERT on specific cyberbullying detection tasks, researchers have demonstrated improvements in cross-platform configurations, particularly when data preprocessing includes considerations for slang and informal language.

Conclusion

The literature review highlights the significance of automated cyberbullying detection in preventing its detrimental effects, including depression and low self-esteem among victims. The studies showed that using models that understand the context of words, like BERT, can lead to better detection of cyberbullying, especially when these models are trained on data that includes online slang. We found that even simpler machine learning approaches can work well, particularly when you combine them with other data points besides just the words themselves. This data could include things like the overall feeling of a message (positive, negative, etc.) or even how the message uses language according to psychology principles. The integration of novel features such as Toxicity features and psycholinguistics features derived from tools like LIWC 2022 and Empath's lexicon enriches the feature space, enhancing model performance. The papers reveal the effectiveness of contextual embeddings, particularly DistilBert, in capturing the suggested meaning of cyberbullying content. Both studies highlighted the need for more data, especially data that reflects how people use language online today, and the importance of developing methods that work for languages other than English. Finally, they both suggested that the best systems might involve a combination of machine learning and human review for the most accurate detection. Overall, by using the best of machine learning, including features beyond just text, and by continuing to improve data collection and human oversight, we can make social media safer for everyone.

References

1. T. H. Teng and K. D. Varathan, "Cyberbullying Detection in Social Networks: A Comparison Between Machine Learning and Transfer Learning Approaches," in IEEE Access, vol. 11, pp. 55533-55560, 2023, doi: 10.1109/ACCESS.2023.3275130.
2. F. Elsafoury, S. Katsigiannis, Z. Pervez and N. Ramzan, "When the Timeline Meets the Pipeline: A Survey on Automated Cyberbullying Detection," in IEEE Access, vol. 9, pp. 103541-103563, 2021, doi: 10.1109/ACCESS.2021.3098979.