Assessing the Efficiency of Machine Learning Algorithms in Detecting Cyberbullying on Twitter

Maram Mohith

2000031507@kluniversity.in

Department of Computer Science and Engineering,

Koneru Lakshmaiah Education foundation, A.P, India

Korukonda Srinivasa Manikanta

2000031813@kluniversity.in

Department of Computer Science and Engineering,

Koneru Lakshmaiah Education foundation, A.P, India

Ravuri Preetham

2000031508@kluniversity.in

Department of Computer Science and Engineering,

Koneru Lakshmaiah Education foundation, A.P, India

Venkata Vara Prasad Padyala

Varaprasad_cse@kluniversity.in

Department of Computer Science and Engineering,

Koneru Lakshmaiah Education foundation, A.P, India

Bommineni Venkata Avinash Chowdary

2000031540@kluniversity.in

Department of Computer Science and Engineering

Koneru Lakshmaiah Education foundation, A.P, India

P Yellamma

pachipala.yamuna@kluniversity.in

Department of Computer Science and Engineering,

Koneru Lakshmaiah Education foundation, A.P, Indi

Abstract

Cyberbullying represents one of the unwanted behaviors which has gotten more and more common since the use of social media, particularly Twitter, has expanded quickly. Cyberbullying can have serious repercussions for its victims. Due to large volumes of data and the intricate structure of online interactions, detecting cyberbullying in real-time is an exhausting task. The purpose of this research proposal is to utilize machine learning techniques that will enhance precision and recall in recognizing instances of cyberbullying on Twitter. In the study, Twitter data is evaluated, trends and patterns are identified, and a wide range of cooperative behaviors are thoroughly assessed. The creation of a framework for collaboration, empirical analysis of patterns of collaboration, evaluation of performance indicators, and improvement of real-time detection capabilities are the main goals.

1.INTRODUCTION

With the rapid growth of social media platforms like Twitter, there has been a corresponding increase in the occurrence of undesirable behaviors among its users (Dalvi et al., 2020). One such behavior is cyberbullying, which can have severe consequences for individuals involved. As a result, it has become imperative to efficiently detect and counteract cyberbullying in real-time, particularly through the analysis of tweets. Because of the large amount of data created and the dynamic nature of online interactions, it is difficult to detect cyberbullying in Twitter data. The intricate patterns and relationships that define cyberbullying may be too complicated for conventional stand-alone methods to completely detect. Collaborative methods that make use of the combined intelligence of several strategies or models are therefore required (Al-garadi et al., 2016).

This research aims to address this need by proposing a collaborative approach for cyberbullying detection on Twitter. The study will empirically evaluate different collaborative patterns and assess their performance in detail. In this research, the main focus will be on enhancing the precision and recall of the cyberbullying detection mechanism based on machine learning algorithms. The research will involve analyzing the bullying dataset, uncovering the trends and patterns, and finally assessing the efficiency of machine learning algorithms. The findings from this study will contribute to the development of more effective and efficient methods for identifying and countering cyberbullying incidents, ultimately promoting a safer and healthier online environment.

1.1 AIMS

This research proposal aims to introduce a collaborative approach for effectively detecting cyberbullying behavior in Twitter data. The proposed framework will combine multiple techniques or models to enhance the accuracy and reliability of detection. By empirically evaluating collaborative patterns and assessing performance metrics such as recall and precision, the research aims to identify the most effective strategies for detecting cyberbullying behavior. The collaborative approach will also focus on enhancing real-time detection capabilities to minimize the impact of cyberbullying incidents.

1.2 OBJECTIVES

To address the growing concern of cyberbullying on Twitter, the objectives of this research proposal are as follows:

1. To develop a collaborative framework that combines multiple techniques or models for the efficient detection of cyberbullying behavior.

2. To empirically evaluate various collaborative patterns in order to identify the most effective strategies for detecting cyberbullying on Twitter.

3. To assess performance metrics, specifically recall and precision, to measure the effectiveness of the collaborative approach compared to stand-alone detection paradigms.

4. To enhance real-time detection capabilities for swift identification and response to cyberbullying incidents on Twitter.

By achieving these objectives, this research proposal aims to improve the accuracy, reliability, and efficiency of cyberbullying detection on Twitter, with the ultimate goal of creating a positive impact on individuals' online experiences.

2. LITERATURE REVIEW

Numerous researchers have investigated the detection of cyberbullying behavior on social media platforms, particularly on Twitter. The following five studies in the same context highlight significant findings and results:

In their study (Muneer & Fati, 2020) developed a machine learning-based approach for cyberbullying detection on Twitter. Their study focused on the analysis of textual and non-textual features, such as the presence of offensive language, user interactions, and user profile characteristics. The findings showed a significant improvement in cyberbullying detection accuracy compared to traditional keyword-based methods and it was found that the logistic regression was outperformed with best f1-score i.e. 92% and best precision was obtained from the SGD i.e. 96%.

In a similar study conducted by (Murshed et al. 2022), the proposed DEA-RNN model was evaluated using a dataset of 10,000 tweets to detect cyberbullying on Twitter. The performance of DEA-RNN was compared to existing approaches like SVM, Bi-LSTM, RNN, MNB, and RF. The results showed that DEA-RNN achieved superior performance with an average accuracy of 90.45% with 89.52% of precision rate. The hybrid model, combining Elman RNN with an optimized DEA for parameter fine-tuning and faster training, demonstrated its effectiveness in detecting cyberbullying incidents on Twitter.

In a different study conducted by (Balakrishnan et al. 2020), the aim was to detect cyberbullying on Twitter by analyzing users' psychological features, including personality traits, sentiment, and emotion. 5,453 tweets pertaining to the hashtag #Gamergate was gathered by the researchers and manually annotated by human specialists, creating a dataset. Using machine learning classifiers like Random Forest and Naïve Bayes, the tweets were divided into four groups: bullies, aggressors, spammers, and normal users. The baseline algorithm included text, user, and network-based features. The findings indicated that incorporating personality traits and sentiment significantly enhanced the detection of cyberbullying incidents, while the influence of emotion was not as pronounced. The research also showed that qualities associated with extraversion, agreeableness, neuroticism, and psychopathy were more significant in predicting the identification of online bullying. By utilizing dimension reduction techniques, the researchers identified ten key features, which were integrated into a single model achieving the highest detection accuracy of 92.88%. The study offers recommendations for applying these findings to effectively address and prevent cyberbullying.

In the same context (Dalvi et al., 2020) addresses the issue of cyberbullying on social media platforms and proposes a machine learning model to automatically detect and prevent bullying actions. Using Naïve Bayes and SVM classifiers for model testing and training, the work focuses on finding word similarities in bullies' tweets. The findings demonstrated that when it came to identifying true positives, both classifiers had accuracy rates of 71.25% for Naive Bayes and 52.70% for SVM. However, on the same dataset, SVM fared better than Naive Bayes, nevertheless. The model was integrated with the Twitter API to fetch tweets and determine if they were classified as bullying or not. This software implementation offers a potential solution for identifying and addressing cyberbullying on Twitter.

Research Gap

Significant progress has been made in the field of cyberbullying detection on Twitter, but there are still notable research gaps that need to be addressed. One such gap is the lack of comprehensive studies on the collaborative approach for cyberbullying detection. While individual techniques and models have been explored, limited research has been conducted on the potential benefits of combining these approaches within a collaborative framework. It is also important to integrate and evaluate various aspects such as textual and non-textual features, linguistic analysis, machine learning algorithms, and social network characteristics in a collaborative manner to identify the most effective strategies. Additionally, there is a need to focus on real-time detection capabilities as most existing studies primarily detect cyberbullying incidents retrospectively, rather than in real-time. By filling these research gaps, the proposed research aims to develop a collaborative framework that effectively detects cyberbullying behavior on Twitter, while also optimizing for real-time implementation and enhancing overall detection capabilities. Ultimately, addressing these gaps will contribute to a more comprehensive and robust approach to combating cyberbullying and promoting online safety.

3. METHODOLOGY

To achieve the aims and objectives of this research proposal, a quantitative methodology will be employed.

The following methods and techniques will be utilized:

1. **Data Collection:** Twitter data containing instances of cyberbullying behavior will be collected using appropriate data scraping techniques. This data will serve as the basis for empirical evaluation and analysis.

2. **Collaborative Framework Development:** A collaborative framework will be developed by integrating multiple techniques or models for cyberbullying detection. This may involve combining machine learning algorithms, natural language processing techniques, and social network analysis methods.

3. **Feature Extraction:** Relevant features will be extracted from the collected Twitter data to represent the characteristics of cyberbullying behavior, such as the presence of offensive language, subjectivity, and sentiment. These features will be quantitatively analyzed to identify patterns associated with cyberbullying incidents.

4. **Empirical Evaluation:** The collaborative patterns within the framework will be empirically evaluated using appropriate statistical methods. Performance metrics, including recall and precision, will be calculated to measure the detection effectiveness of the collaborative approach compared to stand-alone methods.

5. **Comparative Analysis:** The performance of the collaborative approach will be compared to individual detection techniques to determine its strengths and weaknesses. This analysis will provide insights into the added value of collaboration in cyberbullying detection.

6. **Real-Time Implementation:** The collaborative framework will be optimized for real-time implementation to enhance the detection capabilities of cyberbullying incidents on Twitter. This may involve exploring real-time streaming data processing techniques and implementing efficient algorithms to ensure swift identification and response.

7. **Ethical Considerations:** Throughout the research process, ethical considerations will be carefully addressed. Steps will be taken to protect the privacy and confidentiality of individuals involved in the Twitter data, while strictly adhering to data usage guidelines and ethical standards.

The quantitative methodology outlined above will enable the research to systematically develop, evaluate, and optimize a collaborative framework for cyberbullying detection on Twitter, contributing to the overall aim of creating a safer online environment.

4.CONCLUSION

Effective and efficient methods for detecting and avoiding such incidents are required given the increasing rate of cyberbullying on social media services like Twitter. The intention of the current research proposal is to fulfill the immediate gap by offering an integrated approach to identify cyberbullying, with a focus on enhancing precision and recall through using machine learning methods. The targets set out in this research proposal are intended to close significant knowledge gaps and contribute to the creation of more effective methods for detecting and preventing cyberbullying occurrences in real time.

The literature review has shed light on several machine learning-based methods for detecting cyberbullying, emphasizing the importance of language analysis, psychological characteristics, and both textual and non-textual elements in enhancing detection precision.

REFERENCES

- Muneer, A., & Fati, S. M. (2020, October 29). A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter. Future Internet, 12(11), 187. https://doi.org/10.3390/fi12110187

- Murshed, B. A. H., Abawajy, J., Mallappa, S., Saif, M. A. N., & Al-Ariki, H. D. E. (2022). DEA-RNN: A Hybrid Deep Learning Approach for Cyberbullying Detection in

Twitter Social Media Platform. IEEE Access, 10, 25857–25871. https://doi.org/10.1109/access.2022.3153675

- Balakrishnan, V., Khan, S., & Arabnia, H. R. (2020, March). Improving cyberbullying detection using Twitter users' psychological features and machine learning. Computers & Security, 90, 101710. https://doi.org/10.1016/j.cose.2019.101710
- Murshed, B. A. H., Abawajy, J., Mallappa, S., Saif, M. A. N., & Al-Ariki, H. D. E. (2022). DEA-RNN: A Hybrid Deep Learning Approach for Cyberbullying Detection in Twitter Social Media Platform. IEEE Access, 10, 25857–25871. https://doi.org/10.1109/access.2022.3153675
- Dalvi, R. R., Baliram Chavan, S., & Halbe, A. (2020, May). Detecting Twitter Cyberbullying Using Machine Learning. 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS). https://doi.org/10.1109/iciccs48265.2020.9120893
- Al-garadi, M. A., Varathan, K. D., & Ravana, S. D. (2016, October). Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. Computers in Human Behavior, 63, 433–443. https://doi.org/10.1016/j.chb.2016.05.051