Cyberbullying Detection Using Artificial Intelligence: A Survey of Methods and Challenges

Jyothika Sreevalsan*, Meenakshi M*, Swathi K Santhosh*, Viswajith C P*, Rebitha K R†
*Student, Vidya Academy of Science and Technology, Thrissur
†Assistant Professor, Vidya Academy of Science and Technology, Thrissur

***Abstract*—Cyberbullying is a significant issue on social media platforms that is impacting the mental health and well-being of many people and students, as well as their social relationships. With the increasing amount of online communication occurring each day between people and groups, the current methods of moderating and reporting bullying must evolve to become automated so that cyberbullying can be identified and stopped more efficiently.**

This survey presents a summary and organization of the available methods used by AI and ML for the detection and prevention of cyberbullying. Additionally, it analyzes the previously developed studies by reviewing the types of techniques, data, feature extraction methods, optimization strategies and evaluation metrics used by researchers.

The papers in this survey were collected from peer-reviewed publications in major digital libraries. The focus was on recent studies about cyberbullying detection. We analyzed and categorized the chosen studies based on their learning approaches, datasets, and evaluation criteria.

The findings of this review indicate that deep learning and transformer-type models are generally more effective when detecting and preventing cyberbullying because they have improved capabilities of recognizing contextual information and under- standing meaning than traditional machine learning algorithms. Additionally, application of optimization techniques improves performance in identifying bullying by effectively selecting relevant features and tuning hyperparameters.

Although significant progress has been made in our under- standing of the topic, several challenges remain. These challenges include imbalanced datasets, processing of multiple languages and formats and limited ability to explain the rationale for decisions made, as well as ethical issues, and the current lack of real-time implementation of automated systems. Future research should develop integrated explainable, real-time AI- based systems that provide accurate identification of bullying along with proactive measures to prevent it.

***Index Terms*—Cyberbullying Detection, Social Media, Artificial Intelligence, Machine Learning, Deep Learning, Transformer Models, Feature Extraction, Optimization Techniques**

I. INTRODUCTION

The exponential rise of social media has been accompanied by increasing cyberbullying incidences, which have led to negative psychological, social, and academic effects, especially students and young adults. The massive amounts of user, generated content and the changing nature of offensive language make it impossible for human moderators and rule, based systems to be effective.

Consequently, the adoption of artificial intelligence (AI) and machine learning (ML) methods for automated cyberbullying detection and prevention has become a popular trend among researchers.
The transition to deep learning, transformer, based models, and optimization, assisted techniques for better contextual understanding and detection accuracy constitutes the recent work, whereas initial investigations were based on traditional machine learning models. However, issues like multimodal and multilingual data processing, explainability, ethical concerns, and on, the, fly execution are still there despite these advancements.

This survey examines the AI, and ML, based cyberbullying detection methods, which speak to their capabilities, shortcomings, and aspirations for further research.

II. LITERATURE SURVEY

A. Traditional Machine Learning Approaches

Social networking sites have developed an avenue

for bullying through the use of new ways to communicate digitally. The online bullying can take many different forms, so researchers are now creating automated methods to detect and deter different forms of online bullying. [8]

Historically, these automated detection methods utilized traditional machine learning methodologies that rely on pre- classified text data or manually generated text features for the classification of abusive or harmful content. Some of the common machine learning methodologies include Support Vector Machines, Na¨ıve Bayes Classifiers, and TF-IDF feature extraction, and generally, machine learning methodologies performed better than traditional methods, assuming pre- structured datasets exist.

However, while the machine learning methodologies may perform better than traditional machine learning algorithms, the unstructured and disorganized nature of the data obtained from social networking websites remain problematic. [5]

B. Hybrid and Deep Learning–Based Approaches

Current research has looked at ways to improve the accuracy of detecting cyberbullying by integrating deep learning technologies with natural language processing. [8] [5] A number of hybrid machine learning/deep learning approaches have been developed to address issues associated with class imbalance, diversity of language in relation to social media, and the ability to keep up with changing patterns of behaviour. [10]

Ensemble methods together with advanced handling technique such as augmentation and resampling of training datasets have produced more stable results from a variety of different datasets. [10]

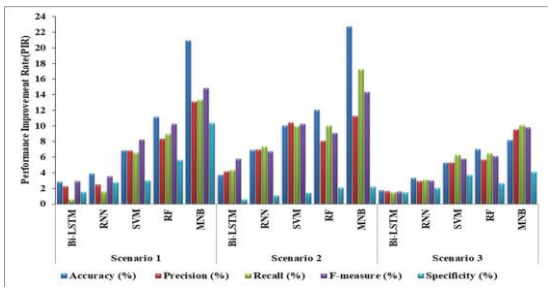

Fig. 1. Performance improvement rate (PIR) (adapted from [3]).

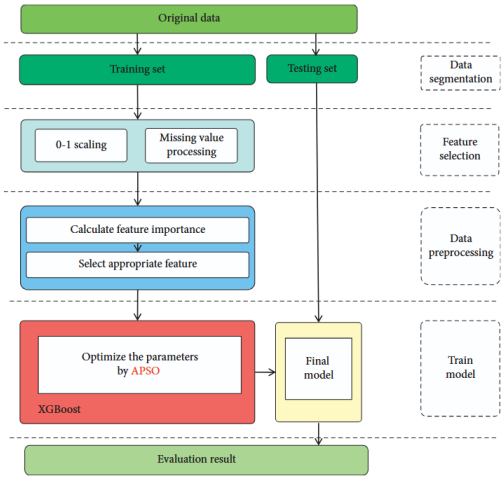

Fig. 2. The process of the APSO-XGBoost credit scoring model(adapted from [9]).

However, most existing deep learning systems use a large amount of computation power, and hence they are typically implemented on servers or through cloud computing systems, resulting in their lack of appropriateness for real-time or mobile solutions. [1]

C. Transformer-Based Cyberbullying Detection Models

As a consequence of their ability to provide an enhanced understanding of the context of a text, transformer-based model architectures have become the most utilized technique for the detection of cyberbullying.

Transformer models (such as BERT, RoBERTa, and Dis- tilBERT) utilize a self-attention mechanism that enables the model to establish long-distance dependencies between data points and identify nuanced differences in the semantic meaning of the words used in a piece of text.

Many previous empirical studies have found dramatic increases in performance, including accuracy, precision, and re- call, as compared to previous model architectures that utilized a convolutional (CNN) or recurrent (RNN) neural network. [5] Despite their capabilities of superior performance, the trans- former models require significant computational resources and memory resources to operate continuously in real-time monitoring, particularly if used on mobile devices with limited resources, unless further optimized to do so. [1]

D. Multimodal Cyberbullying Detection Techniques

The most recent studies on the detection of multimodal cyberbullying indicate that visual content is prevalent in in- stances of cyberbullying. [5] Digital abuse often occurs in connection with visual media – for example, through the use of offensive memes and images, and also screenshots with abusive messages.

A common approach to analyzing this type of content is through convolutional neural networks (CNNs), which can have both high accuracy and low computational complexity when utilized with lighter architecture options, such as Mo- bileNet.

Through the use of multimodal fusion methodologies that combine the text and visual components of the content being analysed, researchers have observed an increase in the detection reliability of multimodal cyberbullying.

In the majority of systems developed for the detection of multimodal cyberbullying to date, there has been a strong emphasis on the classification accuracy of the system, and therefore a failure to address the critical issues of privacy, latency and constraints on the deployment of these systems into real-world conditions. [8]

III. RESEARCH GAP

Despite the advancements made in cyberbullying detection through machine learning/deep learning, there still exists numerous deficiencies in the research currently available. [8]

While most studies focus primarily on enhancing classification accuracy, [1] little thought is given to the practicalities of deploying a system in real-time, scalability of such a system, as well as addressing user privacy concerns, resulting in a limited number of solutions only available in cloud-based or test settings.

Many existing studies use transformer-style architectures which provide excellent context understanding and detection accuracy, however, due to their high computational and memory demands, they cannot be used for continuous on- device monitoring (especially on mobile devices). Currently, multimodal systems (which combine text with visual) provide reliable detection capacity but fail to account for the practical implementation challenges of real-world conditions – such as preserving user privacy/ confidentiality, providing timely detection notifications, and ultimately providing practical use cases. [5]

Ensemble/optimization-based systems have recently enhanced the robustness of many detection systems, most specifically, those used in higher education settings; however, there is little coordination of these methods into single unified, multimodal and real-time cyberbullying detection frameworks, nor have there been proactive approaches developed to prevent cyberbullying incidents by providing proactive intervention to those identified as ”high-risk” users.

Due to these continuing problems, there is a great need for lightweight, privacy-aware, real-time, multimodal-based cyberbullying detection frameworks capable of being deployed on a multitude of devices as well as provide immediate intervention to users displaying at-risk behaviour.

IV. FUTURE RESEARCH DIRECTIONS

In light of the gaps observed in the surveyed works, several promising directions for future research are identified and discussed:

- Lightweight and Mobile-Ready Models: Future cyber- bullying detection systems will take into consideration the need for an efficient model which can apply on both mobile and edge devices, by utilizing various techniques to allows adapting Deep Learning and Transformer-based models to operate at lower resource capabilities, while still maintaining acceptable detection accuracy through methods such as compressing, pruning, quantizing, and Knowledge Distillation.
- Privacy-Preserving Detection Frameworks: Growing demand exists for systems that detect cyberbullying with a focus on privacy preservation and lower dependency on internet-based servers for processing data to create system detections. Instead of processing detection through a remote/cloud-based environment, as is currently done, privacy-aware systems can conduct processing locally within the user’s environment. The following types of technologies support this approach:

on-device inference, federated learning, and secure data aggregation methods.

- Advanced Multimodal Fusion Strategies: There is a need for further research on more durable and reactive strategies for combining different modes of communication (e.g., images, text and contexts). Newer and innovative strategies using the fusion approach or focused multimodal system may create more reliable methods for detecting all types (categories) of social media posts, while providing solutions to issues related to post timing and alignment.
- Explainable and Ethical AI Systems: The implementation of Explainable AI (XAI) in Cyberbullying Detection Models will provide transparency into, and increase interpretability of, prediction outcomes. This increased transparency helps educators, moderators, and end-users comprehend the rationale behind cyberbullying detection outcomes, thus enabling them to make ethical decisions based on this information while also reducing bias in prediction algorithms.
- Proactive Detection and Early Intervention: Instead of classifying a user reactively based on information collected about them after a cyberbullying incident has already occurred, we need to adopt a proactive method- ology to identifying users at risk of being victims of cyberbullying; we also need to develop mechanisms to detect high-risk behaviors associated with cyberbullying. To accomplish this, proactive warning systems, statistical analysis of behavioral trends, and rapid intervention pro- grams are essential to protect the interests of vulnerable populations (e.g., children, students).
- Dataset Standardization and Benchmarking: The in- sufficient standardisation of datasets and assessment processes to allow for common benchmark criteria will continue to hinder progress in this field. Creating publicly available, ethically curated datasets on a large scale with common benchmark methodologies will allow re- searchers and practitioners to compare their results and replicate work more easily as well as help assess how proposed solutions perform in real-world environments.
- Real-Time System Integration and Deployment: Future studies should include full integration of systems from beginning to end; utilizing real-time ingestion of information, rapid inference, efficient and consistent interaction with users. Future studies should have focus placed on the practical limitations of deploying a system like this; such as how to increase energy efficiency while maintaining scalability and continuously upgrading the models used in machine learning.
- Bias Mitigation and Fairness: Models for detecting cyberbullying should not exhibit biased outcomes against any particular group of people based on gender, cultural background, language, or social context. There must be continuous evaluation of bias through techniques related to fairness-aware learning to prevent discriminatory/geographical outcomes.

V. CONCLUSION

This survey presents a review of current advancements in cyberbullying detection using machine learning techniques, Deep learning techniques and NLP (Natural Language Processing) (NLP). While using traditional ML models, their performance tends to be relatively reliable when using well defined feature sets.

However, traditional ML models typically have problems when applying them to complex language patterns. Although deep learning and especially transformer-based systems, no- tably BERT models (e.g., DistilBERT) outperform traditional ML models, they provide an advantage over traditional ML in that they capture the underlying context and semantics of language. Recent literature based on deep learning and BERT models emphasize the significance of using multiple classifiers in an ensemble, addressing the issue of class imbalance and analyzing multiple modalities for robustness.

Furthermore, the hybrid framework utilising deep learning and transformer models may prove to be some of the most effective methods for cyberbullying detection, while future work should concentrate on deploying cyberbullying detection models in real time along with explainability and cross- platform migration to enhance online safety.

REFERENCES

[1] Wachiraporn Tapaopong, Atiphan

Charoenphon, Jakkapong Raksasri, and Taweesak Samanchuen, "Enhancing Cyberbullying Detection on Social Media Using Transformer Models," in Proceedings of the 2024 Technology Innovation Management and Engineering Science Interna- tional Conference (TIMES-iCON), IEEE, 2024.

[2] S. Modha, P. Majumder, T. Mandl, and C. Mandalia, "Detecting and visualizing hate speech in social media: A cyber Watchdog for surveil- lance," Expert Syst. Appl., vol. 161, p. 113725, Dec. 2020

[3] B. A. H. Murshed, J. Abawajy, S. Mallappa, M. A. N. Saif, and H. D. E. Al-Ariki, "DEA-RNN: A Hybrid Deep Learning Approach for Cyberbullying Detection in Twitter Social Media Platform," IEEE Access, vol. 10, pp. 25857–25871, 2022.

[4] T. H. H. Aldhyani, M. H. Al-Adhaileh, and S. N. Alsubari, "Cy- berbullying Identification System Based Deep Learning Algorithms," Electronics, vol. 11, no. 20, p. 3273, Oct. 2022.

[5] Mohammed Mushfiq Ali, Mohammad Nafizul Islam, Saikat Chowdhury, and Md Shafiul Abrar, "Enhancing Cyberbullying Detection: A Compre- hensive Framework Using Machine Learning, Deep Learning, and NLP Techniques," in Proceedings of the 2025 International Conference on Electrical, Computer and Communication Engineering (ECCE), IEEE, 2025.

[6] B. A. Talpur and D. O'Sullivan, "Multi-Class Imbalance in Text Clas- sification: A Feature Engineering Approach to Detect Cyberbullying in Twitter," Informatics, vol. 7, no. 4, p. 52, Nov. 2020.

[7] D. Dess`ı, D. R. Recupero, and H. Sack, "An Assessment of Deep Learning Models and Word Embeddings for Toxicity Detection within Online Textual Comments," Electronics, vol. 10, no. 7, p. 779, Mar. 2021.

[8] Sharmila V. J., Limsa Joshi J., Daya Mary Mathew, Kanishkumar K., Satyavarssheni R. A. V., and Alan Judith A., "Cyberbullying Detection and Prevention in Social Media Using Machine Learning and Deep Learning Techniques," in Proceedings of the 2024 International Confer- ence on Big Data Analytics in Bioinformatics (DABCon), IEEE, 2024

[9] C. Qin, Y. Zhang, F. Bao, C. Zhang, P. Liu, and P. Liu, "XGBoost optimized by adaptive particle swarm optimization for credit scoring," Math. Probl. Eng., vol. 2021.

[10] P. J. Vijayakumar, M. S. Das, P. J. Bansod, S. Kumar, R. Akshaya and S. Praveena, "Detecting and Preventing Cyberbullying Among Higher Education Students on social media with XGBoost and Particle Swarm Optimization," 2025 Global Conference in Emerging Technology (GINOTECH), PUNE, India, 2025.

[11] A. Srivastav, H. V. Reddy and R. Karnati, "A Comparative Study of Machine Learning and Transfer Learning Approaches for Cyber Bullying Detection in Social Networks," 2024 2nd International Conference on Artificial Intelligence Trends and Pattern Recognition (ICAITPR), Hyderabad, India, 2024.