ISSN: 2350-0328

International Journal of AdvancedResearch in Science, Engineering and Technology

Vol. 12, Issue 4, April 2025

Cyberbullying Detection and Prevention Using AI

Manju Sri. S , Mrs. Muthukarupaee .K , Asika .M, Janani .M , Nivethethaa .A.S

U.G. Student, Department of Information Technology, Saranathan College of Engineering, Tiruchirappalli, India
Assistant Professor, Department of Information Technology, Saranathan College of Engineering, Tiruchirappalli, India
U.G. Student, Department of Information Technology, Saranathan College of Engineering, Tiruchirappalli, India
U.G. Student, Department of Information Technology, Saranathan College of Engineering, Tiruchirappalli, India
U.G. Student, Department of Information Technology, Saranathan College of Engineering, Tiruchirappalli, India

ABSTRACT: Cyberbullying has emerged as a significant challenge in the digital era, affecting the mental well-being of individuals, particularly children and teenagers. Traditional detection methods often fail to identify subtle and context-dependent harassment, leading to delayed intervention and prolonged emotional distress for victims. Our approach utilizes XGBoost for text-based classification and InceptionV3 for image-based analysis to effectively detect harmful content. XGBoost, a robust gradient-boosting algorithm, processes textual data, leveraging sentiment analysis and natural language processing (NLP) techniques to classify messages as potentially harmful or non-harmful. In parallel, InceptionV3, a deep learning-based convolutional neural network (CNN), analyses images to identify offensive or inappropriate visual content that may contribute to cyberbullying. By combining sentiment analysis, text classification, and image recognition, our system significantly improves detection accuracy while reducing false positives. This proactive approach fosters a safer online environment by promoting responsible digital behaviour and providing timely alerts to prevent further escalation of harmful interactions. The proposed system has broad applications in social media platforms, educational institutions, and online communities, where cyberbullying remains a pressing concern.

KEY WORDS: Cyberbullying Detection, Machine Learning, Deep Learning, XGBoost, InceptionV3, Natural Language Processing (NLP), Image Classification, Sentiment Analysis, Text Mining, Social Media Safety.

I. INTRODUCTION

In today’s digital age, cyberbullying has become a pervasive issue, significantly impacting mental health, particularly among children and teenagers. As online interactions continue to increase, so does the risk of harassment, which can have severe emotional and psychological consequences for victims. Unlike traditional bullying, cyberbullying occurs in virtual spaces where anonymity and wide audience reach intensify its effects. Identifying and preventing such harmful interactions is crucial to ensuring a safe and respectful online environment. Traditional cyberbullying detection methods rely on predefined keyword matching and rule-based systems, which often struggle to capture the complexities of language, intent, and evolving online behaviour[6]. Subtle forms of harassment, sarcasm, and context-dependent abuse frequently go unnoticed, leading to delayed interventions and prolonged harm. Moreover, the rise of visual content in digital communication necessitates an advanced approach that extends beyond text analysis. To address these challenges, this study introduces an AI-powered cyberbullying detection system that leverages machine learning and deep learning techniques to monitor and analyse online interactions in real-time. The system integrates XGBoost for text classification and InceptionV3 for image-based analysis, ensuring a comprehensive detection mechanism[7]. XGBoost, an efficient gradient-boosting algorithm, classifies textual content by analysing sentiment and linguistic patterns, identifying potentially harmful messages. Simultaneously, InceptionV3, a deep convolutional neural network, detects inappropriate or offensive visual content contributing to cyberbullying. By combining sentiment analysis, text classification, and image recognition, the proposed system enhances accuracy, reduces false positives, and enables timely intervention.

Cyberbullying On Social Media Sites

The major contributors to cyberbullying are social networking sites. The dynamic nature of these sites helps in the growth of online aggressive behaviour. The anonymous feature of user profiles increases the complexity to identify the bully.

Social media is popular due to its connectivity in the form of networks. But this can be harmful when rumours or bullying posts are spread into the network which cannot be easily controlled. Twitter and Facebook can be taken as examples which are popular among various social media sites. According to Facebook users have more than 150 billion connections which gives the idea about how bullying content can be spread within the network in a fraction of time. To manually identify these bullying messages over this huge network is difficult. There should be an automated system where such kinds of things can be detected automatically thereby taking appropriate action[1]. Researches have shown that cyberbullying has negative effects on victims. The victim mainly consists of women and teenagers. Incentive effect on mental and physical health of the victims. Ease in such kind of activities higher is the risk of depression leading to suicidal cases. Therefore, to control cyberbullying there is need of automatic detection or monitoring systems.

Detection Models For Cyberbullying

Research has shown that machine learning can be used to identify cyber bullying efficiently. We tried to develop a real time cyber bullying monitoring system which will identify bullying and non-bullying tweets. Every time a user selects this option real time tweets will be obtained and classification of these tweets will happen. Along with this analysis of the tweets will be displayed showing the severity and frequency of bullying. Frequency of bullying means how many times a user has bullied others on twitter and severity explains how many bullying words are present in the tweet. These features can be used to take disciplinary actions against such users and can provide analysis for the extent of bullying taking place on twitter. A report will also be displayed showing the results here the date of tweet, the tweet, the twitter handle of the user will be displayed[4].

II. EXISITING SYSTEM

The existing system leverages machine learning and natural language processing (NLP) to automatically analyse user-generated content on social media. The core objective is to classify and identify harmful or abusive language that may indicate cyberbullying. The paper explores several deep learning models, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks. These models are trained on labelled datasets containing bullying and non-bullying text data. The system utilizes pre-processing techniques such as tokenization and stop-word removal to clean and prepare the data for training. The study compares the accuracy and performance of various models to determine the most effective architecture for cyberbullying detection[9]. This existing system demonstrates the potential of AI-driven solutions to improve user safety on digital platforms by automatically flagging abusive behaviour in real-time.

Disadvantages

- Requires high computational resources for training and inference.
- Long training times with diminishing returns after multiple epochs.
- Struggles to understand complex or long-range contextual language.
- Needs large volumes of labelled data, which is costly to prepare.
- Models may not generalize well across different social media platforms.

III. PROPOSED SYSTEM

XGBoost (Extreme Gradient Boosting) is a powerful machine learning algorithm used to classify text messages as cyberbullying or non-cyberbullying by following a structured process. It begins with data preprocessing, where text is cleaned by removing special characters, stop words, and irrelevant symbols, followed by tokenization and stemming for normalization. Next, feature extraction techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) and word embeddings are applied to convert text into numerical features. The model is then trained using labelled datasets that include examples of both cyberbullying and non-cyberbullying content, allowing it to learn patterns based on linguistic cues, sentiment, and context. Once trained, the model can predict whether new incoming messages contain cyberbullying. For image analysis, InceptionV3, a deep convolutional neural network (CNN), is employed. The process

ISSN: 2350-0328
International Journal of AdvancedResearch in Science, Engineering and Technology
Vol. 12, Issue 4, April 2025

involves image preprocessing through resizing and normalization, followed by feature extraction to identify offensive gestures, hate symbols, or explicit content. Finally, the model classifies images as harmful or non-harmful using a SoftMax activation function.

Advantages

- Achieves high accuracy in both text and image classification.
- Enables multimodal detection by analysing both text and images.
- XGBoost ensures fast and efficient text processing.
- InceptionV3 effectively captures complex visual features.
- Scalable across platforms and adaptable to various datasets.
- Suitable for real-time monitoring and detection.
- Combines TF-IDF and word embeddings for better text context understanding.
- Supports automated content moderation with minimal human intervention.

IV. SYSTEM ARCHITECTURE

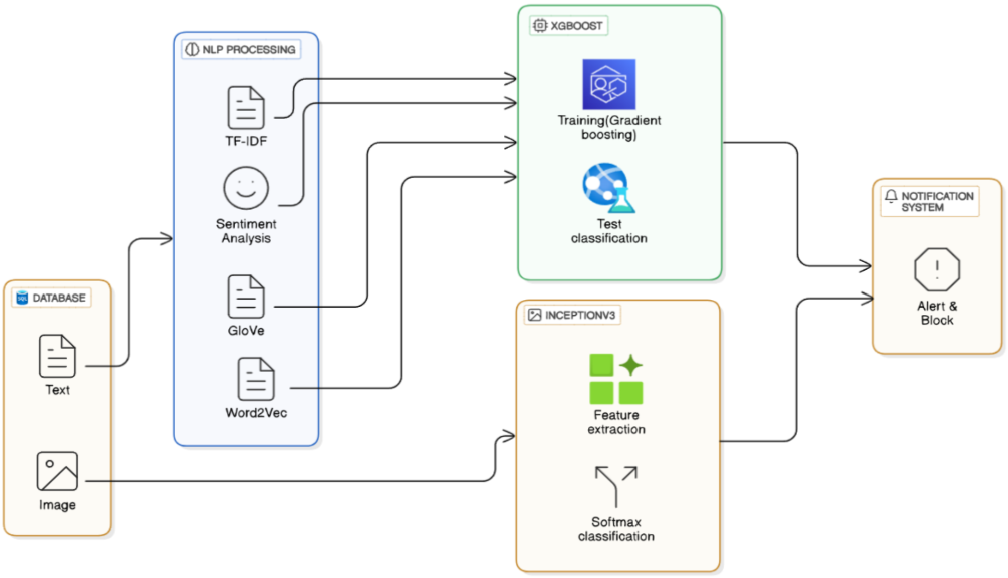

Fig1. System Architecture

The cyberbullying detection system is composed of multiple modules, each handling a specific aspect of content analysis. These modules work together to detect harmful text and images using machine learning and deep learning techniques. The integration of these components ensures accurate, real-time identification and alerting of cyberbullying incidents.

Text-Based Cyberbullying Detection Module

This module is responsible for detecting cyberbullying in textual content such as comments, posts, or messages. It is powered by XGBoost (Extreme Gradient Boosting), a powerful machine learning algorithm known for its high accuracy and efficiency. The module begins by preprocessing the textual data—removing special characters, URLs, and stop words, followed by tokenization and stemming. Text is then transformed into numerical features using TF-IDF and word embeddings, enabling the model to understand contextual and semantic patterns. XGBoost functions by building an ensemble of decision trees, where each subsequent tree focuses on minimizing the errors of the previous ones using gradient boosting. To avoid overfitting, it uses regularization techniques like tree pruning and shrinkage. This module contributes to the system's ability to flag harmful textual content in real time.

Image-Based Cyberbullying Detection Module (InceptionV3)

This module is dedicated to identifying harmful visual content, such as offensive gestures, hate symbols, or explicit imagery, using the deep learning model InceptionV3. Through a deep stack of convolutional and pooling layers, the model captures varying levels of visual detail. It then reduces data dimensionality using Global Average Pooling (GAP) and classifies content using fully connected layers and a SoftMax activation function, which outputs probabilities indicating whether an image is harmful or safe. This module enhances the system's capability to detect and prevent visual forms of cyberbullying.

Real-Time Detection and Alert Module

This module integrates the outputs of both the text and image detection models into a real-time monitoring system. Whenever a new post or image is uploaded, the respective modules analyse the content. If cyberbullying is detected in either text or image, this module triggers immediate alerts to platform moderators, parents, or educators. The module ensures swift action by providing detailed logs and visual indicators of the flagged content, facilitating quick responses and maintaining a safe online environment.

V. EXPERIMENTAL RESULTS

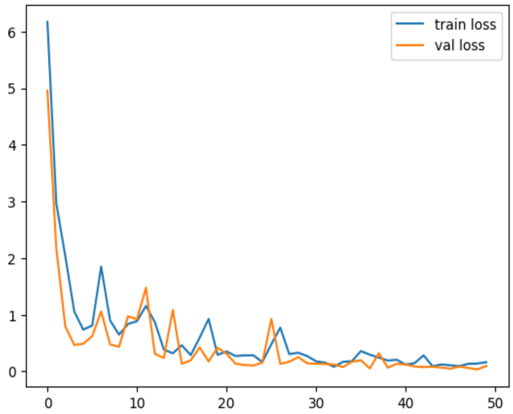

Fig 2. Training and Validation Loss

This graph depicts the training and validation loss across 50 epochs. The loss values for both the training and validation sets decrease rapidly in the initial stages, indicating that the model is learning effectively. The curves eventually stabilize near zero, suggesting that the model has converged well. The minimal gap between training and validation loss indicates that overfitting is minimal and the model maintains consistency on unseen data.

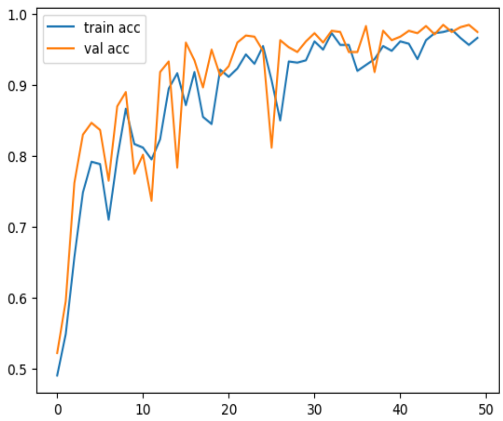

Fig 3. Training and Validation Accuracy

This graph presents the training and validation accuracy over 50 epochs. The model's accuracy improves significantly within the first few epochs and continues to rise, ultimately achieving over 90% accuracy for both training and validation datasets. The consistency between the training and validation accuracy curves reflects strong generalization capabilities, implying that the model can accurately detect cyberbullying in both seen and unseen data.

VI. CONCLUSION AND FUTURE WORK

Cyberbullying remains a significant issue in the digital era, affecting individuals' mental and emotional well-being, particularly among children and teenagers. Traditional detection methods often fall short in identifying subtle and context-dependent harassment, leading to delayed intervention and prolonged distress for victims. To address this challenge, the proposed AI-driven cyberbullying detection system integrates XGBoost for text classification and InceptionV3 for image-based analysis, ensuring a comprehensive and efficient approach to detecting harmful online interactions. By leveraging machine learning and deep learning, this system significantly enhances the accuracy of cyberbullying detection while minimizing false positives. XGBoost processes textual content using sentiment analysis and feature extraction, allowing for the classification of harmful messages with greater precision. Meanwhile, InceptionV3 analyses images, identifying offensive or inappropriate content that contributes to cyberbullying. The combination of these techniques enables real-time monitoring of online interactions, ensuring prompt detection and response to cyberbullying incidents. The proposed system not only fosters a safer digital environment but also empowers parents, educators, and platform moderators to take immediate action against online harassment. By providing timely alerts and interventions, this AI-driven solution helps protect vulnerable users from the harmful effects of cyberbullying, promoting responsible digital behaviour and awareness.

Future enhancements to the proposed AI-driven cyberbullying detection system could involve integrating multimodal analysis by incorporating audio and video content alongside text and image inputs, enabling the system to better capture context and tone in complex interactions. Advancements in natural language processing can enhance the system's ability to understand sarcasm, slang, and evolving online language, while explainable AI models can offer transparency behind decisions, fostering user trust. Real-time response systems can be improved through edge computing to enable faster detection and intervention. Incorporating user-specific behavioural patterns and psychological profiling may allow for more personalized detection strategies, increasing sensitivity to at-risk individuals. To ensure robustness, the system should also be trained to resist adversarial attacks and bias, supported by continuous learning from real-world data. Collaborations with mental health professionals and ethicists can guide the responsible use of AI in this space, ensuring privacy, fairness, and accountability. By embracing these enhancements, future systems can provide even more effective, inclusive, and proactive solutions to mitigate the harm caused by cyberbullying.

REFERENCES

[1] Al-Marghilani, Abdulsamad. "Artificial Intelligence-Enabled Cyberbullying-Free Online Social Networks in Smart Cities." International Journal of Computational Intelligence Systems, vol. 15, no. 9, 2022

[2] Asfia Sabahath, Arshiya Begum, Pundru Chandra Shaker Reddy, Marepalli Radha, Jay Pawar, Mithra C, "A hybrid framework for image cyberbullying recognition using transfer deep learning," Proceedings of the 2024 Global Conference on Communications and Information Technologies (GCCIT), pp. 1–7, IEEE, Oct. 2024.

[3] Belal Abdullah Hezam Murshed, Jemal Abawajy , Suresha Mallappa, Mufeed Ahmed Naji Saif, Hasib Daowd Esmail Al-ArikiA, "DEA-RNN: Hybrid deep learning approach for cyberbullying detection in Twitter social media platform," in Proceedings of the 2022 IEEE International Conference on Data Science and Computer Application (ICDSCA), pp. 1–6, IEEE, 2022.

[4] Gomez, Christopher E., Marcelo O. Sztainberg, and Rachel E. Trana, "Curating Cyberbullying Datasets: A Human–AI Collaborative Approach." International Journal of Bullying Prevention, vol. 4, no. 1, pp. 35–46, 2022.

[5] J. Sathya and F. M. H. Fernandez, "Enhancing cyberbullying detection in social media: Leveraging ontology with Skip-gram optimization," Proceedings of the 9th International Conference on Communication and Electronics Systems (ICCES), pp. 927–934, IEEE, 2024.

[6] Manuel F. López-Vizcaíno, Francisco J. Nóvoa, Victor Carneiro, and Fidel Cacheda, "Early detection of cyberbullying on social media networks," in Proceedings of the 2022 IEEE International Conference on Big Data (Big Data), pp. 1234–1241, IEEE, Dec 2022.

[7] Milosevic, Tijana, Kathleen Van Royen, and Brian Davis, "Artificial Intelligence to Address Cyberbullying, Harassment and Abuse: New Directions in the Midst of Complexity." International Journal of Bullying Prevention, vol. 4, no. 1, pp. 1–5, 2022.

[8] M. Dadvar and K. Eckert, "Cyberbullying detection in social networks using deep learning based models; a reproducibility study," arXiv preprint arXiv:1812.08046, Dec 2018.

[9] M. H. Obaida, S. M. Elkaffas, and S. K. Guirguis, “Deep learning algorithms for cyber-bullying detection in social media platforms,” *IEEE Access*, vol. 12, pp. 76901–76908, IEEE, 2024.

[10] M. Saravanan Karthikeyan, D. Abiatha Kumari, S. Murali, K. Paul Joshua, R. Santhanam Krishnan, J. Relin Francis Raj, "Towards safer digital spaces: Automated detection of cyberbullying through multi-modal learning," Proceedings of the International Conference on Sustainable Communication Networks and Application (ICSCNA), pp. 480–489, IEEE, May 2024.