

School of Computer Science Engineering and Information Systems

Winter Semester 2024-2025

Department of Information Technology

BITE498J – Project – II

Review -1

**Cyber Bullying Detection System**

**21BIT0705  
Smaran Lal Shrestha**

**Under the Guidance of**

Prof. BHAVANI S

**Guide Signature with date**

**ABSTRACT:**

Cyberbullying is an escalating issue in the digital world, and addressing it requires advanced detection systems to create safer online spaces. This project focuses on developing a comprehensive solution to detect cyberbullying through both text and image analysis. For text-based detection, a hybrid approach is used, combining rule-based filtering techniques such as predefined word lists and regex patterns to capture variations of offensive language. Additionally, transformer-based models like BERT are fine-tuned using cyberbullying datasets to understand context and detect subtle nuances. Data augmentation methods, including paraphrasing and misspelling variations, are incorporated to enhance the model's accuracy and robustness.

For image-based detection, the system employs computer vision techniques to analyze harmful visual content, such as memes or inappropriate images. Pre-trained deep learning models, like Convolutional Neural Networks (CNNs) or Vision Transformers (ViTs), are fine-tuned to identify harmful content effectively. The system integrates both text and image detection into a unified pipeline, providing a scalable solution for online platforms. The goal of this project is to help mitigate the psychological impact of cyberbullying and foster a safer and healthier online environment.

**Keywords:** Cyber Bullying, Natural Language Processing (NLP), Image Recognition, Customized Alerts

**INTRODUCTION:**

Cyberbullying has become a widespread and pressing issue in the digital age, with the rise of social media, online communication platforms, and virtual interactions. It poses significant threats to individuals' mental health, self-esteem, and overall well-being. Detecting and mitigating cyberbullying is a challenging task due to the nuanced and context-dependent nature of harmful behavior, which can manifest through text, images, or even subtle patterns like sarcasm and slang.

Traditional detection methods often fail to capture the complexity of online interactions, where language varies across cultures, and visual content can carry implicit meanings. This project aims to address these challenges by developing a comprehensive cyberbullying detection system that leverages advanced Natural Language Processing (NLP), image recognition, and multilingual support to identify harmful content effectively. By integrating these technologies into a scalable and efficient pipeline, the system aspires to provide real-time detection, empowering users and moderators to create safer digital environments.

**PROBLEM STATEMENT:**

Cyberbullying has become a pervasive issue in the digital world, posing significant threats to individuals' mental health and well-being. Detecting and addressing cyberbullying is challenging due to its nuanced nature, with harmful content often disguised through slang, sarcasm, abbreviations, or images. Existing methods are either limited in scope, fail to capture contextual subtleties, or lack the ability to analyze multimodal content such as text and images simultaneously.

The absence of scalable, efficient, and context-aware detection systems leaves online platforms struggling to mitigate the psychological and social impacts of cyberbullying. This project aims to address these challenges by leveraging advanced machine learning and deep learning techniques to develop a comprehensive system capable of accurately detecting cyberbullying and promoting safer digital spaces.

**OBJECTIVES:**

The objective of our Cyberbullying Detection System is to develop a model that can accurately detect cyberbullying in online platforms using both text and image analysis. The goals are to:

1. Detect harmful text using NLP models like BERT, classifying content as abusive or non-abusive.
2. Analyze images for offensive content using computer vision techniques.
3. Support multiple languages to detect cyberbullying globally.
4. Provide real-time detection and timely alerts.
5. Generate customized alerts for users and moderators based on detected content.

This system aims to create safer online environments by efficiently identifying and addressing cyberbullying in both text and visual content.

**SCOPE OF THE PROJECT:**

The primary goal of this project is to develop a system for detecting cyberbullying across online platforms using text and image analysis. This research aims to explore the effectiveness of different machine learning techniques in accurately identifying cyberbullying content. By employing advanced Natural Language Processing (NLP) models and image recognition methods, the study will compare their performance in detecting harmful behavior in both text and images. Despite the growing advancements in AI, detecting cyberbullying remains a complex challenge due to the subtle and evolving nature of abusive language and visual content. The scope of this project is to contribute to the development of more accurate, real-time, and context-aware systems that can enhance online safety and provide actionable insights for users and moderators.

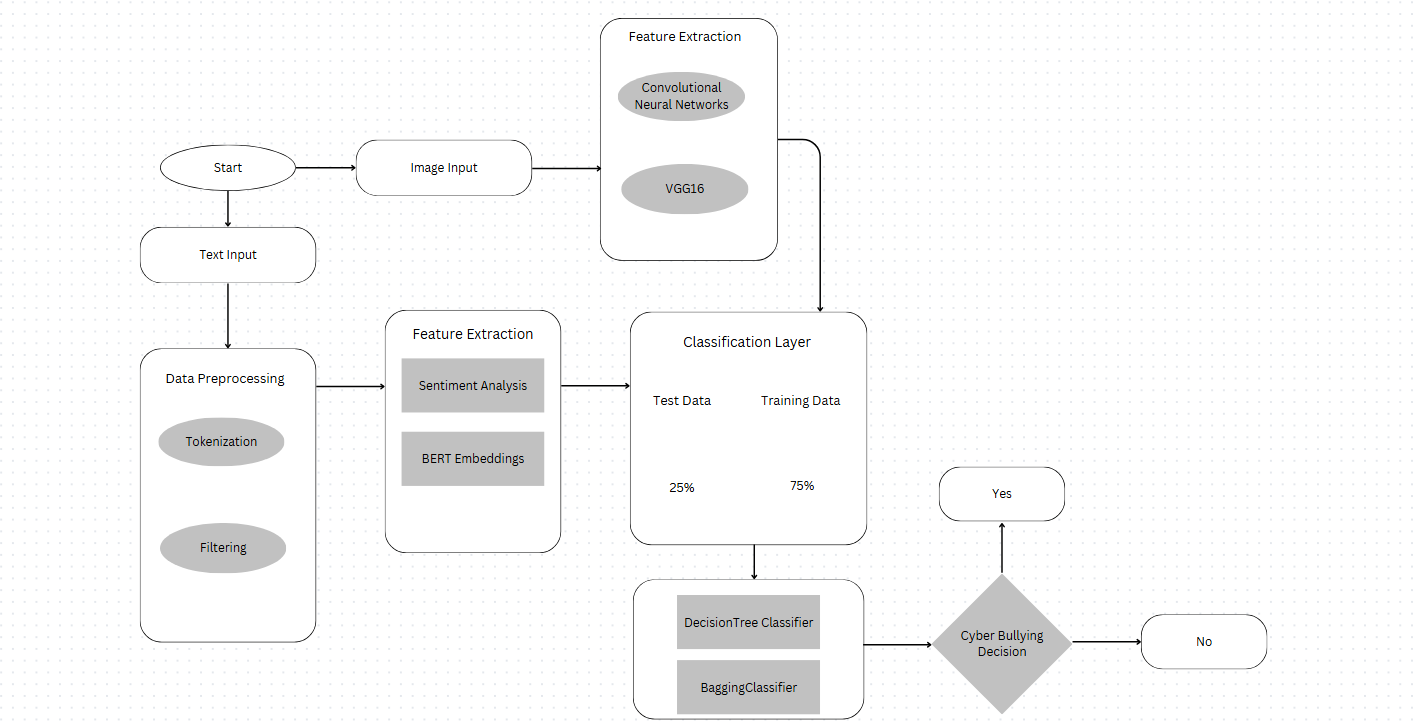
**PROPOSED SYSTEM:**

The proposed system aims to detect cyberbullying in real-time across both text and image content on online platforms using advanced machine learning techniques. It will utilize transformer-based models like BERT to identify harmful language in text and Convolutional Neural Networks (CNNs) for detecting offensive visuals, such as memes or inappropriate images. The system will also support multiple languages, making it applicable in diverse cultural contexts. By providing real-time detection and customized alerts, the system will allow immediate action to be taken by users, moderators, or parents, promoting a safer and more supportive online environment.

**LITERATURE SURVEY:**

|  |  |  |  |
| --- | --- | --- | --- |
| **S.NO** | **TITLE** | **MERITS** | **DEMERITS** |
| 1 | DEA-RNN: A Hybrid Deep Learning Approach for Cyberbullying Detection in Twitter Social Media Platform | Combines Elman type RNN and Dolphin Echolocation Algorithm (DEA) for fine-tuning parameters and reducing training time. Outperforms existing models in accuracy, precision, recall, F1-score, and specificity. Effective for short text classification in social media. | The hybrid model’s complexity could be difficult for real-time implementation due to its dependence on fine-tuning parameters. The model's performance might be dataset-dependent and may not generalize well across diverse data sources. |
| 2 | Cyber Bullying Detection Using Machine Learning | The paper proposes a new representation learning method, SMSDA, which enhances the stacked denoising auto-encoder with semantic dropout noise and sparsity constraints to improve the detection of cyberbullying. It focuses on robust numerical representation learning for text messages, which is crucial for effective detection in social media. | The method might require further real-world validation and might struggle with the diversity of social media language, such as sarcasm, emojis, or indirect bullying. |
| 3 | A Comparative Study of Machine Learning Approaches for Cyberbullying Detection in Digital Forums | he paper evaluates multiple machine learning models (LR, SVM, RF, DT, NB, XG) for detecting cyberbullying and identifies Random Forest as the most effective model. It uses two datasets (10,000 and 20,000 comments) and explores tokenization, stop words, and Bag of Words for text preprocessing. | The system may face challenges with algorithm bias, data privacy concerns, and may require improvements to handle more diverse or complex forms of bullying. |
| 4 | Bilingual Cyber Bullying Detection System: Enhancing Online Safety | The system supports both English and Hindi (transliterated into English) for detecting cyberbullying. It uses advanced data extraction and preprocessing techniques, as well as machine learning algorithms. A robust web application is created using Python, Streamlit, Pandas, and Scikit-learn for real-time text analysis. | The system may face challenges with algorithm bias, data privacy concerns, and may require improvements to handle more diverse or complex forms of bullying. |
| 5 | Enhancing Cyber Bullying Detection Using Convolutional Neural Network | The paper demonstrates the application of deep learning (CNN) and NLP to detect cyberbullying in social media, specifically focusing on Twitter. It uses Twitter API to gather data and applies CNN to classify tweets as bullying or non-bullying. | The paper might face challenges like dataset quality, false positives, and the complexity of detecting subtle forms of bullying, especially in short, ambiguous text |
| 6 | Cyber Bullying and Toxicity Detection Using Machine Learning | The paper explores a novel approach for detecting  Cyberbullying and toxicity with machine learning algorithms such as SVM, Logistic Regression, Naive Bayes, KNN, and Random Forest, achieving 90% accuracy for text and 84.5% for images using MobileNetV2. | The paper might have limitations due to the dataset size and potential challenges in addressing all forms of cyberbullying or toxicity, especially in multilingual or mixed-content scenarios. |
| 7 | Automated Multimodal Detection and Reporting of Cyber Bullying using ML and Encryption | Uses machine learning techniques for detecting cyberbullying in text, photos, and videos with high accuracy (98.5%, 88.7%, and 80.7%). Implements  cryptographic techniques (RSA, SHA-256, AES) for data security and integrity. Aims to streamline reporting, creating a safer digital environment. | May face challenges in real-time detection across all types of media and require continuous updates to adapt to new forms of cyberbullying. |
| 8 | Cyberbullying Detection using Deep Learning Models in Bengali Language | The study focuses on detecting cyberbullying in the Bengali language, a low-resource language. It uses deep learning models like Bi-LSTM and CNN, and the CNN model outperforms the Bi-LSTM model in terms of precision, recall, and F1 score. The research uses a public dataset from Kaggle. | The study is limited to the Bengali language, and its application may not be generalized to other languages. |
| 9. | Cyber-Bullying Detection Via Text Mining and Machine Learning | This paper proposes the use of machine learning and natural language processing (NLP) for detecting cyberbullying in real-time before content is uploaded. The research highlights how real-time detection can reduce the circulation of harmful messages. | Limitations of current solutions and inefficient algorithms in detecting cyberbullying. It doesn't offer a concrete framework for deployment in all online environments. |

**SYSTEM ARCHITECTURE:**



**REFERENCES:**

1. Siddhartha, K., Kumar, K. R., Varma, K. J., Amogh, M., & Samson, M. (2022). Cyber bullying detection using Machine learning. 2022 2nd Asian Conference on Innovation in Technology (ASIANCON), 1–4. https://doi.org/10.1109/asiancon55314.2022.9909201
2. Murshed, B. a. H., Abawajy, J., Mallappa, S., Saif, M. a. N., & Al-Ariki, H. D. E. (2022b). 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>
3. Rohini, D. S., & Ramchander, M. (2023b). A Comparative Study of Machine Learning Approaches for Cyber bullying Detection in Digital Forums. ICAICCIT, 332–338. https://doi.org/10.1109/icaiccit60255.2023.10466143
4. Uke, S., Amrutkar, K., Kulkarni, A., Kasliwal, T., & Konde, A. (2024). Bilingual Cyber Bullying Detection System: Enhancing Online Safety. 2024 8th International Conference on I-SMAC, 837–842. https://doi.org/10.1109/i-smac61858.2024.10714645
5. Jadhav, R., Agarwal, N., Shevate, S., Sawakare, C., Parakh, P., & Khandare, S. (2023). Cyber Bullying and Toxicity Detection Using Machine Learning. 2023 3rd International Conference on Pervasive Computing and Social Networking (ICPCSN). https://doi.org/10.1109/icpcsn58827.2023.00017
6. Deekshitha, M., Kesavamoorthy, R., Shah, P., & Upadhyay, A. U. (2024). Automated Multimodal Detection and Reporting of Cyber Bullying using ML and Encryption. ," 2024 International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS), 1198–1205. https://doi.org/10.1109/icicnis64247.2024.10823159
7. Beniwal, R., Jha, S., Mehta, S., & Dhiman, R. (2023). Cyberbullying Detection using Deep Learning Models in Bengali Language. 2023 3rd International Conference on Intelligent Technologies (CONIT), 1–5. https://doi.org/10.1109/conit59222.2023.10205775
8. Sharma, C., Ramakrishnan, R., Pendse, A., Chimurkar, P., & Talele, K. T. (2021). Cyber-Bullying detection via text mining and machine learning. 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT), 1–6. https://doi.org/10.1109/icccnt51525.2021.9579625
9. Vadnagarwala, F. H., & Agrawal, A. J. (2024). Aggression Detection in Social Media Texts using Machine Learning and Deep Learning Models. 2024 OPJU International Technology Conference (OTCON), 1–5. https://doi.org/10.1109/otcon60325.2024.10687824
10. Behzadi, M., Harris, I. G., & Derakhshan, A. (2021). Rapid Cyber-bullying detection method using Compact BERT Models. 2021 IEEE 15th International Conference on Semantic Computing (ICSC). https://doi.org/10.1109/icsc50631.2021.00042