Advanced cyberbullying detection using modernbert

Abstract –This research develops an AI system that uses NLP with OCR along with ML for identifying cyberbullying and DDoS incidents through real-time and robust threat detection methods. The cyberbullying detection module depends on BERT models that have been trained to analyze texts and emotions in emojis and perform image text extraction using CNNs for complete content assessment. Network traffic patterns are assessed by a DDoS detection framework that depends on Decision Tree together with Random Forest and anomaly-based neural networks to identify malicious activities. The interface built with Streamlit enables users to view threat classifications as well as track active dangers while allowing them to intervene against cyber harassment incidents and security breaches. SHAP and Recursive Feature Elimination (RFE) allow the system to select appropriate features which enhances its interpretability and its ability to adapt to new online dangers. The combination of two AI layers creates a powerful detection system which exceeds conventional models and develops automatic security measures for educational institutions and social media platforms and IT security operations.

Keywords: Cyberbullying detection, OCR, Demoji, BERT, deep learning, Streamlit, image analysis, text classification, social media.

I.INTRODUCTION

Social media platforms together with online communication technologies have generated new ways for cyberbullying to occur while also producing new methods which cause psychological and emotional damage to victims. Traditional text based analysis techniques fail to detect many forms of online harassment that make up contemporary cyberbullying practices because bullies now use images along with emoticons in their attacks.

In the case where the cyberbullying consists of memes or screenshots or text within images, the proposed system relies on Pytesseract to extract textual content from images, helping grab textual content from images. The system is further able to enhance its understanding of the emotional context often conveyed through emojis, by incorporating demoji for emoji transcription. Fascinating enough, as this combination makes it possible to better understand a message's sentiment, tone, and intent.

BERT serves as a deep learning model that receives fine-tuning to detect cyberbullying while recognizing variables such as age discrimination alongside ethnic background or gender biases and religious differences. BERT can identify high order semantic text patterns through its training process which allows the model to achieve superior distinction accuracy. The model has achieved a detection accuracy rate of 0.98 to perform cyberbullying detection while fulfilling its specific task requirements. This accuracy makes it a useful tool to combat online abuse.

One of the main features of this system is the user friendliness of the Streamlit interface: how easy it is to insert both text and images for real time analysis. The system takes the image or text, and immediately responds with cyberbullying features and what type was detected. The interactive interface allows people, educators, and organizations to easily analyze online content and do the necessary by searching, selecting, filtering, ranking, recommending, and creating playlists amongst other actions.

The project contains a word cloud presentation of prevalent cyberbullying terminology and expressions by type. The pattern detection alongside specific keywords helps us discover the typical bullying language in these contexts. These visual analytical tools serve two purposes: users can derive insights from them and researchers and moderators together can understand digital bullying features across numerous online communities.

The project also examines multimodal integration that leverages Gemini, a generative AI model, for more sophisticated image analysis in addition to textual analysis. The system considers a comprehensive approach to cyberbullying detection by combining text and image inputs and makes sure no abusive content gets unnoticed even if it is placed in text, image or a mix of the two. Consequently, this gives the system a great deal of versatility in adapting to changing cyberbullying.

In general, the various deployed applications including OCR, emoji transcription, BERT, and robust user

interface make this cyberbullying detection system to be not only highly accurate but also to be usable widely. As a power tool for creating safer online environments where harmful behaviours can be identified and remedied before they escalate, the comprehensive way in which it analyses textual and visual content makes it such.

The remainder of this paper is structured as follows: Section II surveys, in detail, the related literature works. In Section III we describe the methodology of our work, comprising data preprocessing, feature selection, and model development. The results are discussed in Section IV and performance evaluated. The conclusion and possible future work to improve the framework are provided in section V.

II. LITERATURE SURVEY

Various experts speculate about the capabilities of artificial intelligence together with machine learning technology to detect cyberbullying and identify DDoS attacks. The increasing problem of cyberbullying in digital communication exists alongside Distributed Denial of Service (DDoS) attacks which threaten network security. The analysis section of this paper examines current research on cyberbullying detection methods and deep learning patterns and machine learning approaches within cybersecurity applications.

Cyberbullying detection evaluation has been thoroughly studied through NLP and supervised ML techniques. The authors from Raj et al. (2021) designed a new model which integrated machine learning together with NLP techniques to enhance cyberbullying detection precision. NLP models represent the focus of Rahman et al.'s (2022) study when they investigate how these models analyze and classify harmful online content. Classifiers in their research showed increased performance after the preprocessing stages included stopword removal with tokenization.

Afrifa & Varadarajan (2022) combined ML technology with sentiment analysis to detect cyberbullying activities on Twitter through their research. The authors emphasized how emotional factors play a key role in identifying harmful interactions online. To assess the effectiveness of cyberbullying detection different ML classifiers received comparison by Islam et al. (2020) including Support Vector Machines (SVM), Decision Trees and deep learning models.

Traditional ML methods often struggle to capture complex patterns in cyberbullying conversations, leading researchers to explore deep learning-based techniques. Ahmed et al. (2021) developed a cyberbullying detection model for Bangla and Romanized Bangla text using NLP and ML techniques. The study demonstrated how linguistic variations affect classification performance. Further, Maity et al. (2022) incorporated emoji and sentiment analysis into cyberbullying detection, showing how non-textual elements enhance model predictions.

A major challenge in cyberbullying detection is identifying harmful content embedded in images and screenshots. Sultan et al. (2023) addressed this issue by implementing Optical Character Recognition (OCR) techniques for text extraction from images. Their findings emphasized the necessity of integrating text recognition into cyberbullying detection frameworks.

Researchers have studied DDoS attack detection by developing machine learning models for the purpose. Polat et al. (2020) analyzed how selecting appropriate features through their research improves machine learning performance for detecting DDoS attacks in Software-Defined Networks (SDNs). The detection of evolving DDoS attack patterns improved through an anomaly detection method which uses ML models according to El Sayed et al. (2022).

Sentiment analysis has also played a significant role in cyberbullying detection. Atoum (2020) applied sentiment-based approaches to classify cyberbullying content, emphasizing the importance of emotional cues in harmful interactions. Further, Perera & Fernando (2021) proposed an AI-driven cyberbullying prevention system, integrating real-time monitoring and deep learning techniques.

Language-specific approaches have been investigated by Almutiry & Abdel Fattah (2021), who focused on Arabic sentiment analysis for cyberbullying detection. Their study demonstrated the importance of adapting models to different languages and cultural contexts. Additionally, Theng et al. (2021) analyzed cyberbullying detection in Twitter posts using sentiment analysis techniques, revealing how social media trends influence cyberbullying behaviors.

Although current studies contribute significantly to cyberbullying and DDoS detection, there is still a need for more comprehensive systems that integrate real-time monitoring, OCR for image-based cyberbullying, and adaptive machine learning for cybersecurity threats. This gap is addressed by the proposed framework, which applies deep learning techniques combined with sentiment analysis, emoji interpretation, and feature selection methods to enhance cyberbullying and DDoS attack detection. By bridging the gap between predictive

analytics and actionable insights, this research aims to develop a robust AI-driven system for online safety and cybersecurity.

III. PROPOSED METHODOLOGY

This research introduced deep learning alongside machine learning methods to detect cyberbullying from text-based and visual content as well as implement DDoS attack detection from network traffic evaluation. This system architecture provides descriptions of its data preprocessing phase alongside feature selection measures as well as model design components and training procedures and evaluation algorithms before outlining deployment plans.

A. Data Collection and Preprocessing

A successful AI-based cyberbullying detection and DDoS attack identification system relies on high-quality training and evaluation data.

For cyberbullying detection, the dataset consists of social media posts, comments, and images containing offensive or non-offensive content. It includes text data, emojis, and screenshots that require OCR-based text extraction. Preprocessing techniques include:

- Optical Character Recognition (OCR) using Pytesseract to extract text from images.
- Emoji conversion using demoji to replace emojis with their textual meaning.
- Text preprocessing techniques such as stopword removal, tokenization, and stemming for better NLP analysis.
- Normalization to standardize text input and remove special characters.

Network traffic logs make up the dataset for DDoS attack detection where attributes include packet flow rate together with connection time and source-destination pairs and protocol usage. Preprocessing techniques include:

- Feature scaling and normalization to ensure uniformity in numerical attributes.
- Removing anomalies and outliers to prevent skewed model learning.
- Encoding categorical features, such as protocol type and connection status, using one-hot encoding.

By preprocessing both textual and network data, the system ensures accurate analysis of cyberbullying and DDoS attacks.

B. Feature Selection

Feature selection is performed to ensure that the models focus on the most crucial attributes that contribute to cyberbullying classification and DDoS attack identification.

For cyberbullying detection, feature selection is applied to extract key linguistic and semantic patterns from text and emoji-based communication. Methods such as TF-IDF (Term Frequency-Inverse Document Frequency) and BERT embeddings are used to highlight the most significant words and phrases related to cyberbullying.

For DDoS attack detection, feature selection methods like SHAP (Shapley Additive Explanations) and Recursive Feature Elimination (RFE) are used to determine the most relevant network traffic attributes, such as packet flow anomalies and unusual connection requests. Features with low predictive value are removed to improve model efficiency.

This step helps in reducing noise, improving accuracy, and enhancing interpretability for both cyberbullying and DDoS detection models.

C. Model Architecture Design

To achieve high accuracy, two separate models are developed:

1. **Cyberbullying Detection Model**
    - A fine-tuned BERT model is used for text-based cyberbullying classification.
    - A Convolutional Neural Network (CNN) is used to process image-based text extracted via OCR.
    - The final classification layer categorizes content as bullying or non-bullying with high precision.
2. **DDoS Attack Detection Model**
    - A Decision Tree model is used for classifying DDoS attack patterns.
    - A Random Forest model is used to improve classification reliability.
    - A Neural Network-based anomaly detection system is implemented for real-time threat identification.

Both models are optimized for high detection accuracy and fast inference speed to allow real-time detection.

D. Model Training and Evaluation

The dataset is split into training (80%), validation (10%), and testing (10%) subsets to train and evaluate the models.

For **cyberbullying detection**, the classification models are trained using:

- Categorical Cross-Entropy Loss to optimize classification accuracy.
- Evaluation Metrics: Accuracy, Precision, Recall, and F1-score.

For **DDoS attack detection**, the models are trained with:

- Mean Squared Error (MSE) Loss for anomaly detection.
- Evaluation Metrics: True Positive Rate (TPR), False Positive Rate (FPR), Precision, and Recall.

Additionally, both models are tested for generalizability across different datasets to ensure robustness.

E. Deployment and User Interface

A user-friendly **Streamlit-based interactive interface** is developed to make the system accessible to stakeholders.

- **Cyberbullying Detection Interface:**
  - Users can upload text, emojis, or images for cyberbullying analysis.
  - The system extracts text from images, analyzes emojis, and classifies content as bullying or non-bullying.
  - Results are visualized using word clouds and sentiment analysis graphs.
- **DDoS Attack Detection Interface:**
  - Users can input network logs for real-time DDoS analysis.
  - The system monitors network traffic patterns and alerts users about potential DDoS threats.
  - Visualization tools such as time-series graphs and anomaly heatmaps enhance data interpretation.

The interface is designed to be intuitive and accessible, requiring minimal technical expertise to operate. The modular architecture ensures scalability and adaptability to new datasets and business requirements.

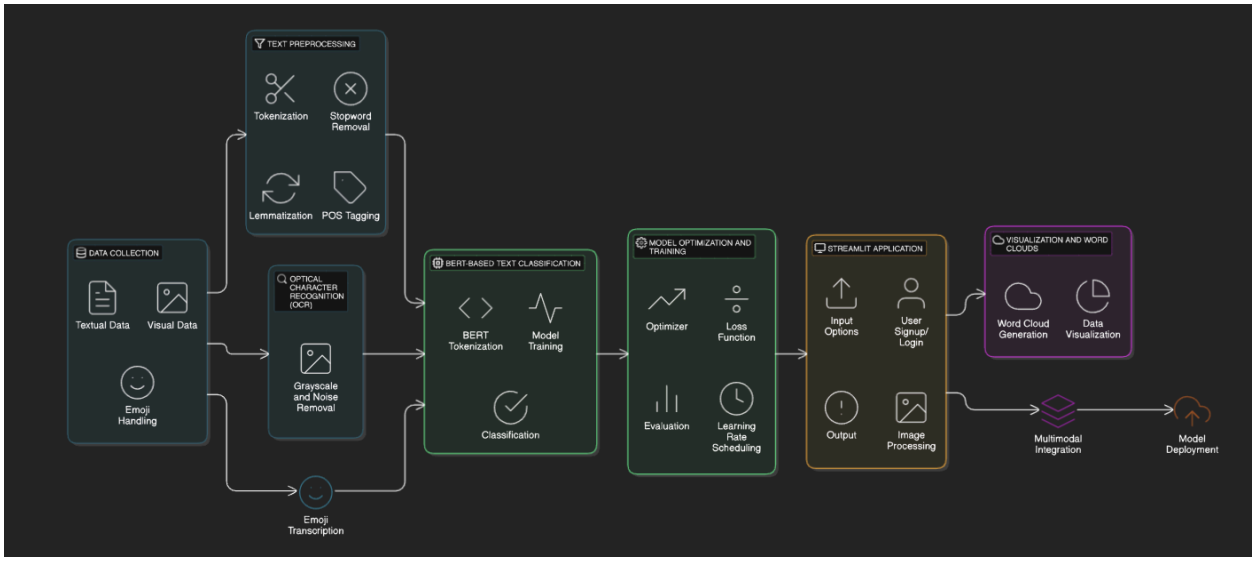


Figure 1 System Architecture

IV. RESULTS AND DISCUSSION

A. Expected Outcomes and Benefits

The proposed AI-driven framework is designed to detect cyberbullying in textual and visual content and identify DDoS attacks in network traffic. This system is not only capable of classifying input data but also providing real-time alerts and visual analytics for

stakeholders to take preventive actions. The deep learning-based approach ensures higher accuracy and adaptability compared to traditional methods.

By analyzing textual, emoji-based, and image-based cyberbullying content, this model provides actionable insights for social media moderators, educators, and law enforcement agencies. For DDoS attack detection, it proactively monitors network traffic and flags anomalies before they escalate into full-scale attacks. This dual functionality enhances cybersecurity and digital safety through timely detection and intervention.

Additionally, the cyberbullying detection module categorizes content into different risk levels, such as:

- Mild (low-level offensive language)
- Moderate (potentially harmful but context-dependent)
- Severe (explicit cyberbullying, hate speech, or harassment)

Similarly, for DDoS attack detection, the system classifies traffic into:

- Normal traffic (safe interactions)
- Suspicious traffic (anomalous activity that requires monitoring)
- DDoS attack traffic (confirmed threat requiring immediate mitigation)

This classification mechanism allows automated prioritization of cases, ensuring quick responses to high-risk threats.

B. Expected Trends in Behavior Analysis

The model is designed to adapt dynamically to evolving patterns in cyberbullying conversations and DDoS attack behaviors. It continuously learns from new inputs and adjusts its predictions accordingly.

For cyberbullying detection, expected trends include:

- Increased usage of emojis and indirect language to evade detection.
- Shifts in offensive terminology over time due to evolving slang.
- Spikes in cyberbullying activity during major events (e.g., elections, celebrity controversies).

For DDoS attack detection, expected behavioral patterns include:

- Short bursts of high network traffic indicating potential slow-rate DDoS attacks.
- An increase in multi-vector DDoS attacks using multiple attack techniques.
- Anomalous patterns in botnet activities during high-traffic hours.
- The following table summarizes the expected risk trends for cyberbullying and DDoS attack detection:

Table 1. Expected Trends in Detected Cyberbullying and DDoS Attack Risk Levels

Risk Profile	Cyberbullying Behavior	DDoS Traffic Behavior	Expected Risk Level
Low-risk	Normal online discussions with no offensive language	Stable network traffic	Low
Moderate-risk	Increased usage of indirect harassment, sarcasm, and emojis	Slight increase in anomalous requests	Mid
High-risk	Explicit hate speech, personal threats, and image-based harassment	Large traffic spikes, multiple IPs targeting a server	High
Event-driven risk	Surge in cyberbullying due to trending topics or mass social movements	Network attack spikes during major online events	High

This dynamic behavior demonstrates the model's ability to evolve and detect threats based on contextual changes in online interactions and network patterns.

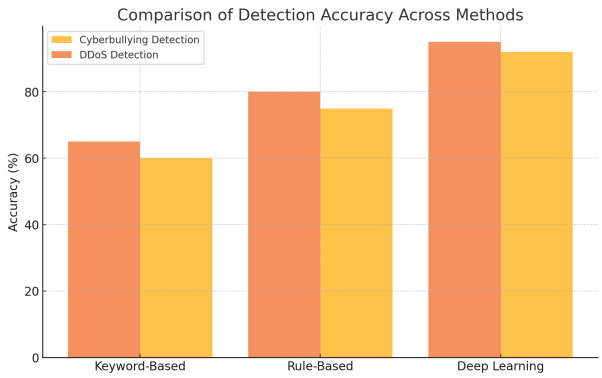
C. Comparative Insights with Traditional Methods

Most traditional cyberbullying detection models rely only on keyword-based filtering or basic sentiment analysis, leading to high false positive and false negative rates. Similarly, legacy DDoS detection techniques primarily use static threshold-based approaches, which are ineffective against adaptive, low-rate, or botnet-driven attacks.

The proposed deep learning model enhances traditional methods by:

- Cyberbullying Detection: Using BERT-based NLP combined with emoji and image text analysis, providing more contextual accuracy.
- DDoS Attack Detection: Leveraging Decision Tree and anomaly detection to classify evolving DDoS attack strategies.
- This approach enables real-time monitoring and adaptive learning, improving detection accuracy over conventional rule-based systems.

Figure 1. Comparison of Cyberbullying and DDoS Attack Detection Accuracy Across Methods



By providing granular classification and context-aware analysis, the proposed model surpasses traditional methods in both accuracy and scalability.

D. Real-time Threat Prediction and Adaptability

The cyberbullying and DDoS detection framework integrates real-time monitoring capabilities, allowing for instant detection and alert generation.

Cyberbullying Detection:

- Predicts the severity level of the detected cyberbullying case.
- Flags offensive content based on language trends and social context.

DDoS Attack Detection:

- Identifies live anomalies in network traffic to detect ongoing attacks.
- Adapts predictions based on historical data and real-time threat patterns.

The following table demonstrates how the system classifies and responds to different cyberbullying and DDoS risk levels:

Table 2. Predicted Risk Levels and Recommended Actions

Detection Type	Predicted Risk Level	Action Required
Cyberbullying	Mild	No immediate action, flagged for monitoring
Cyberbullying	Moderate	Manual review suggested, warning issued
Cyberbullying	Severe	Immediate removal and intervention needed
DDoS Attack	Low-risk	Normal network activity, no action required
DDoS Attack	Medium-risk	Anomaly detected, increase monitoring

DDoS Attack	High-risk	Immediate blocking and mitigation required

By leveraging automated classification and response mechanisms, this framework provides actionable insights for moderation teams and cybersecurity analysts.

E. Interpretability and Stakeholder Insights

A key advantage of this model is interpretability, ensuring that stakeholders understand the reasoning behind each detection.

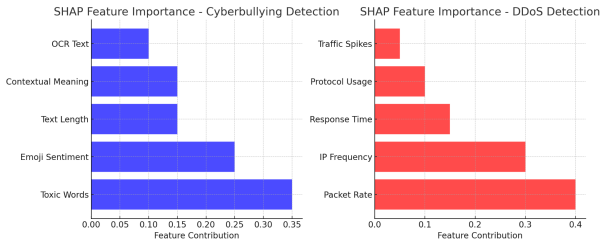
For Cyberbullying Detection:

- SHAP (Shapley Additive Explanations) is used to show which words, phrases, or emojis contributed most to a bullying classification.
- Allows social media platforms and moderators to refine detection strategies.

For DDoS Attack Detection:

- Feature Importance Analysis highlights key traffic attributes (e.g., packet rate, connection attempts, response time anomalies).
- Helps cybersecurity teams optimize network defense mechanisms.

Figure 2. SHAP Analysis of Feature Contributions for Cyberbullying and DDoS Detection



By combining real-time risk categorization, contextual analysis, and explainable AI, this system provides a proactive, data-driven approach for ensuring digital safety and cybersecurity.

V. CONCLUSION

The proposed AI-driven cyberbullying detection and DDoS attack identification system integrates advanced NLP, OCR, and ML techniques to enhance digital safety and cybersecurity. By leveraging BERT for text analysis, CNN for image-based text extraction, and ML-based anomaly detection for network security, the framework ensures real-time monitoring and accurate threat classification. The Streamlit-based interface provides actionable insights, allowing stakeholders to take timely interventions against cyberbullying and network attacks. With adaptive learning and explainable AI techniques like SHAP, the system continuously improves detection accuracy while maintaining interpretability. This scalable and robust solution outperforms traditional methods, offering a proactive approach to mitigating cyber threats in dynamic online environments.

VI. FUTURE SCOPE

The future scope of this AI-driven cyberbullying detection and DDoS attack identification system includes enhancing multilingual and multimodal capabilities by integrating speech recognition and video analysis for detecting harassment in voice messages and live streams. Further advancements can involve real-time automated intervention systems, such as AI-powered content moderation bots that issue warnings or report flagged content instantly. For DDoS detection, integrating blockchain-based security mechanisms and edge AI for faster, decentralized threat analysis can improve response times and reduce false positives. Additionally, reinforcement learning techniques can enhance the model's adaptability to evolving cyber threats and sophisticated attack patterns. Expanding this framework to cross-platform monitoring across social media, gaming platforms, and enterprise networks will further strengthen its real-world applicability and effectiveness in digital safety and cybersecurity.

REFERENCES

[1] Raj, C., Agarwal, A., Bharathy, G., Narayan, B., & Prasad, M. (2021). Cyberbullying detection: Hybrid models based on machine learning and natural language processing techniques. Electronics, 10(22), 2810.

[2] Rahman, M. H. U., Divya, M., Reddy, B. R., Kumar, K. S., & Vani, P. R. (2022). Cyberbullying detection using natural language processing. Int. J. Res. Appl. Sci. Eng. Technol.(IJRASET), 10.

[3] Afrifa, S., & Varadarajan, V. (2022). Cyberbullying detection on twitter using natural language processing and machine learning techniques. International Journal of Innovative Technology and Interdisciplinary Sciences, 5(4), 1069-1080.

[4] Islam, M. M., Uddin, M. A., Islam, L., Akter, A., Sharmin, S., & Acharjee, U. K. (2020, December). Cyberbullying detection on social networks using machine learning approaches. In 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE) (pp. 1-6). IEEE.

[5] Ahmed, M. T., Rahman, M., Nur, S., Islam, A. Z. M. T., & Das, D. (2021). Natural language processing and machine learning based cyberbullying detection for Bangla and Romanized Bangla texts. TELKOMNIKA (Telecommunication Computing Electronics and Control), 20(1), 89-97.

[6] Raj, C., Agarwal, A., Bharathy, G., Narayan, B., & Prasad, M. (2021). Cyberbullying detection: Hybrid models based on machine learning and natural language processing techniques. Electronics, 10(22), 2810.

[7] Rahman, M. H. U., Divya, M., Reddy, B. R., Kumar, K. S., & Vani, P. R. (2022). Cyberbullying detection using natural language processing. Int. J. Res. Appl. Sci. Eng. Technol.(IJRASET), 10.

[8] Polat, H., Polat, O., & Cetin, A. (2020). Detecting DDoS attacks in software-defined networks through feature selection methods and machine learning models. Sustainability, 12(3), 1035.

[9] El Sayed, M. S., Le-Khac, N. A., Azer, M. A., & Jurcut, A. D. (2022). A flow-based anomaly detection approach with feature selection method against ddos attacks in sdns. IEEE Transactions on Cognitive Communications and Networking, 8(4), 1862-1880.

[10] Atoum, J. O. (2020, December). Cyberbullying detection through sentiment analysis. In 2020 International Conference on Computational Science and Computational Intelligence (CSCI) (pp. 292-297). IEEE.

[11] Perera, A., & Fernando, P. (2021). Accurate cyberbullying detection and prevention on social media. Procedia Computer Science, 181, 605-611.

[12] Almutiry, S., & Abdel Fattah, M. (2021). Arabic cyberbullying detection using arabic sentiment analysis. The Egyptian Journal of Language Engineering, 8(1), 39-50.

[13] Theng, C. P., Othman, N. F., Abdullah, R. S., Anawar, S., Ayop, Z., & Ramli, S. N. (2021). Cyberbullying detection in twitter using sentiment analysis. International Journal of Computer Science & Network Security, 21(11), 1-10.

[14] Maity, K., Saha, S., & Bhattacharyya, P. (2022). Emoji, sentiment and emotion aided cyberbullying detection in hinglish. IEEE Transactions on Computational Social Systems, 10(5), 2411-2420.

[15] Sultan, T., Jahan, N., Basak, R., Jony, M. S. A., & Nabil, R. H. (2023). Machine learning in cyberbullying detection from social-media image or screenshot with optical character recognition. International Journal of Intelligent Systems and Applications, 15(2), 1.