A Comprehensive Study on the Effectiveness of Machine Learning Models to Mitigate the Impact of Cyberbullying

Shweta Singh, Fred Barez, and Riti Gour
Department of Aviation and Technology
San Jose State University, San Jose, California 95192, USA

Email:{shweta.singh01, fred.barez and riti.gour}@sjsu.edu

***Abstract*—Cyberbullying affects 20-40% of digitally connected adolescents, yet existing automated detection systems achieve highly variable accuracy and lack ethical decision-making mechanisms. This study presents a unified AI framework combining Machine Learning (ML), Deep Learning (DL), and hybrid models to address cyberbullying detection across four major platforms: Wikipedia Talk, Twitter, Facebook, and YouTube. The framework introduces a novel confidence-based moderation layer that converts prediction certainty into three actionable tiers: automatic blocking (high confidence), user warnings (moderate confidence), and content publishing (low confidence). By addressing persistent challenges in platform adaptability, model interpretability, and ethical decision-making, the proposed framework provides a scalable and robust solution for real-time cyberbullying detection that strikes a balance between high technical performance and responsible, real-world deployment.**

***Index Terms*—Cyberbullying, Convolutional Neural Network.**

I. Introduction

The exponential growth of social media platforms, such as Wikipedia Talk, X (formerly Twitter), Meta (formerly Facebook), and YouTube, has transformed how people communicate, share information, and express their opinions online [1]. However, this expansion has also boosted the prevalence of cyberbullying, deliberate, harmful behavior conducted via digital means, which poses serious psychological, emotional, and social consequences for affected users, especially among younger users and vulnerable communities.

As user-generated content scales, manual moderation becomes increasingly impractical, and there is a compelling demand for automated, scalable, and trustworthy cyberbullying detection systems. Traditional detection approaches often rely on keyword matching or bag-of-words representations that struggle to represent the subtle, implicit, and contextualized dynamics of online abuse. Such systems typically struggle with context, sarcasm, or coded languages, as well as emotionally laden dialogue that does not contain explicitly harmful words. To address these challenges, recent research has turned to more sophisticated Natural Language Processing (NLP) techniques and Artificial Intelligence (AI)-driven models [2].

This study evaluates the effectiveness of machine learning, deep learning, and hybrid models for cyberbullying detection across four major social media platforms. In addition to model comparisons, the work introduces a confidence-based moderation layer that simulates ethical decision-making in real-world scenarios. This mitigation system translates model confidence scores into three actions: block, warn, or publish, offering a fair and calibrated alternative to binary enforcement logic and helping to reduce false positives (shown in Fig.1). Given the real-world impact of moderation decisions on user trust and safety, such systems must adhere to the principles of fairness, transparency, and privacy. This framework, therefore, embeds Responsible AI practices at its core, ensuring that decisions are not only accurate but also ethically sound.

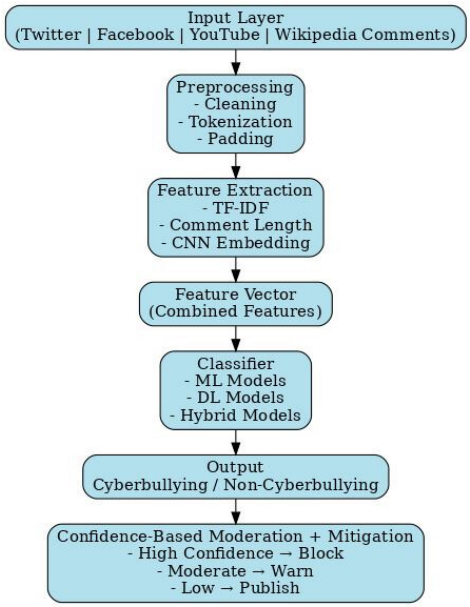

Fig. 1. System Architecture

The rest of the paper is organized as follows: Section II reviews the related literature. Section III presents the model architecture and cross-platform application for cyberbullying detection. Section IV presents the performance evaluation, and Section V concludes with future research directions.

II. Background Literature

Early cyberbullying detection systems relied heavily on rule-based techniques such as keyword filtering and manually crafted heuristics. While easy to implement, these methods often fail to capture sarcasm, contextual abuse, or evolving

slang, resulting in poor accuracy and high false-positive rates. As the limitations of rule-based systems became evident, researchers began to explore data-driven approaches using Artificial Intelligence (AI).

The introduction of Machine Learning (ML) marked a critical shift in cyberbullying detection. Classical classifiers such as Logistic Regression, Random Forest, and XGBoost, when paired with engineered features like TF-IDF scores, sentiment polarity, and comment length, demonstrated significant performance improvements [1], [3]. However, these models still struggled with complex contextual and sequential dependencies in conversations [4]. Deep Learning (DL) models further advanced the field by enabling automated feature extraction and semantic pattern recognition. Convolutional Neural Networks (CNNs) were among the first to achieve high performance in offensive content detection by learning local n-gram patterns [2], [5].

The challenge of getting both accuracy and practicality led researchers to combine traditional and deep learning methods. Hybrid models have been shown to outperform standalone approaches, particularly when interpretability and real-time constraints are essential [6]. One such framework, known as ProTect, combines CNN and RNN layers with proactive detection capabilities [7]. This model not only achieved high accuracy but also offered mitigation pathways for early intervention.

A. Platform-Specific Dynamics

Despite improvements in model architecture, many studies remain limited to a single social media platform, resulting in poor generalization. For instance, in [8], the authors developed a hybrid DL model that performed well on Twitter data but lacked evaluation across structurally different platforms like Wikipedia or YouTube. Each platform exhibits unique linguistic characteristics—Twitter enforces brevity, leading to compact or coded messages; Facebook allows extended harassment through comment threads; YouTube comments are often multimodal or timestamped; and Wikipedia Talk pages foster formal yet emotionally charged collaborative discussions.

Deep Learning-based detection has been explored across multiple platforms and highlighted that platform-specific models, while accurate in isolation, often fail when deployed in cross-domain scenarios [9]. The need for session-based analysis has also been emphasized, as single-comment classifiers often overlook the escalation and recurrence of bullying behavior over time [10].

B. Ethical Challenges and Confidence-Based Moderation

While technical advances have improved detection accuracy, ethical moderation remains a key challenge. Many detection models are "black boxes" lacking transparency and fairness in decision-making [7]. A confidence-based approach to moderation, such as the SafeText system, employs thresholding to automate "block" or "allow" actions based on model certainty; however, it does not account for ambiguous cases that may require human oversight [11].

It has also been highlighted that neural networks often suffer from poor calibration, with confidence scores failing to accurately reflect the true reliability of predictions [12]. It leads to overconfident false positives, which undermine user trust and the integrity of moderation systems. This growing awareness has led to demand for moderation systems that are not only technically sound but also ethically transparent, allowing content to be blocked, warned, or published based on well-calibrated prediction confidence.

C. Research Gaps and Motivation

Despite considerable progress, several key limitations remain in the current body of research. First, most models are developed and tested on a single platform, leading to poor generalizability and limited cross-platform validation. Second, there is a lack of systematic benchmarking studies that compare Machine Learning, Deep Learning, and hybrid models side by side on the same datasets. Finally, most moderation systems operate on binary logic, flagging or allowing content without offering graduated response tiers or human-in-the-loop escalation pathways, which limits their fairness and adaptability.

This research fills these gaps by developing a comprehensive AI framework that works across multiple platforms, comparing traditional machine learning, deep learning, and hybrid approaches on data from Wikipedia, Twitter, Facebook, and YouTube. The study introduces a confidence-based moderation approach that translates prediction certainty into three distinct actions: blocking, warning, or publishing content, providing a more nuanced and scalable alternative to simple yes/no decisions.

III. Model Architecture for Cyberbullying Detection

A. Overview and Role of Machine Learning Models

This section outlines the theoretical foundations of each machine learning model employed in this study. It describes its functional role in detecting cyberbullying across four major platforms: Wikipedia Talk, X (formerly Twitter), Meta (Facebook), and YouTube. These datasets differ in terms of language complexity, comment length, tone, and frequency of cyberbullying, allowing the models to be tested for generalization across diverse digital environments.

A key input to many of these models was a TF-IDF vector— a numerical representation of each comment based on Term Frequency–Inverse Document Frequency. This technique emphasizes words that are important in a specific comment but rare across the entire dataset, enabling models to focus on significant, context-relevant vocabulary for cyberbullying detection. These vectors were often combined with other engineered features, such as sentiment polarity and comment length, to enrich the input data.

We worked on six different machine learning approaches. The Logistic Regression (LR) method was applied to TF-IDF vectors, sentiment scores, and comment metadata to classify a comment. Its high interpretability and low computational

cost made it especially useful for quick, transparent moderation decisions—ideal for platforms like X and YouTube. On Wikipedia and Facebook, it served as a robust baseline model even with more nuanced or extended comment styles.

The Support Vector Classifier (SVC) was used across all four datasets to handle high-dimensional, sparse text features. It effectively detected borderline or ambiguous comments, particularly on Facebook and Wikipedia, where indirect language often obscured harmful intent. Random Forest (RF) identified key cyberbullying indicators—like profanity, extreme sentiment, and comment length—performing well on noisy YouTube data and offering strong feature insights for structured platforms like Wikipedia.

Gradient Boosting (GB) was used to detect subtle aggression, sarcasm, and coded language—especially on Facebook and Wikipedia—and effectively surfaced less obvious bullying patterns in YouTube and Twitter data. XGBoost, a regularized and scalable version of GB, handled large datasets from Wikipedia Talk and X efficiently while maintaining accuracy on noisy YouTube comments. Naive Bayes (NB) achieved high recall for aggressive keywords in imbalanced Facebook and YouTube data and proved useful for rapid, exploratory model evaluation across all datasets.

B. Overview and Role of Deep Learning Models

This section elaborates on the deep learning architectures employed in this study for cyberbullying detection. These models were selected for their ability to learn from sequential unstructured text data. In the implementation pipeline, all deep learning models shared a standard preprocessing structure, which included text cleaning, tokenization, padding, and embedding.

First, we apply a Convolutional Neural Network (CNN) for its ability to capture local patterns in text, such as phrases or n-grams that indicate abusive behavior. Each user comment was tokenized, padded to a fixed length, and then passed through an embedding layer before being entered into the CNN architecture. CNN performed exceptionally well on platforms like YouTube and Twitter, where cyberbullying often occurs in short and repetitive phrases.

Next, Long Short-Term Memory (LSTM) networks were applied to model long-range dependencies within sentences. In contrast to CNN, LSTM considers the sequential structure of the input, making it well-suited for platforms like Wikipedia and Facebook, where users express hostile intent through longer or more syntactically complex sentences. The input sequences were processed through an embedding layer and passed to LSTM cells. LSTM was instrumental in capturing hidden aggression that builds over a sequence, such as threats embedded in sarcastic or narrative formats.

Gated Recurrent Units (GRU) offered a lighter, faster alternative to LSTM with comparable performance. GRUs simplify the gating mechanism, resulting in fewer parameters and faster training. The input data followed the same embedding and padding pipeline as with the LSTM. GRUs were particularly useful for detecting emotional bursts or offensive sentiment in short-form content on platforms like Twitter and YouTube, where speed and interpretability are prioritized.

Additionally, Bidirectional LSTM (BiLSTM) extended the LSTM by processing text in both forward and backward directions. This dual perspective enabled the model to understand both the preceding and following context of a word or phrase—an essential capability for detecting sarcasm, implied hostility, or double meanings. For instance, a comment like "Oh, you're such a genius" could be interpreted as either praise or sarcasm depending on the surrounding words. BiLSTM was highly effective on Facebook and Wikipedia, where users often express complex emotions and nuanced hostility.

C. Overview and Role of Hybrid Models

Hybrid models in this study were designed to combine the complementary strengths of machine learning (ML) and deep learning (DL) to enhance cyberbullying detection across diverse platforms. The hybridization strategies were implemented in two forms: (1) Feature-level fusion, where shallow features (e.g., TF-IDF vectors, sentiment polarity, comment length) are concatenated with deep representations (e.g., CNN outputs). (2) Model-level stacking: where the output of one model serves as input to another, forming a layered decision-making pipeline.

We now discuss the hybrid models in detail. In CNN + Logistic Regression (Feature-Level Hybrid), CNN-generated embeddings with traditional features like TF-IDF, sentiment polarity, and comment length were then passed the fused features to a Logistic Regression classifier. The model achieved strong and interpretable performance, especially on Wikipedia, where long-form discussions required both contextual and statistical understanding.

CNN + BiLSTM, a model-level hybrid approach, was used to classify user comments as cyberbullying or not by combining local and contextual text features. CNN layers extracted shallow phrase-level patterns, which were passed to a bidirectional LSTM to capture contextual dependencies in both forward and backward directions. The output was a binary label: 1 (cyberbullying) or 0 (non-cyberbullying). This model was chosen for its ability to detect nuanced language and emotional tone, making it highly effective on platforms like Wikipedia and X, where comments often carry layered meanings.

Next, the CNN + LSTM (Model-Level Hybrid) approach combined a CNN for local pattern extraction and an LSTM for capturing long-term dependencies in text. It received embedded and padded user comments as input and produced a binary output: 1 (indicating cyberbullying) or 0 (indicating non-cyberbullying). This architecture was beneficial on Wikipedia and Meta, where longer comments required detecting threats or evolving aggression over time. In CNN + GRU (Model-Level Hybrid) approach, a CNN is used for extracting local abusive patterns and a GRU for modeling short-term dependencies in text sequences. GRU offered faster training and fewer parameters than LSTM, making it ideal for shorter, emotionally charged comments on platforms like YouTube and X.

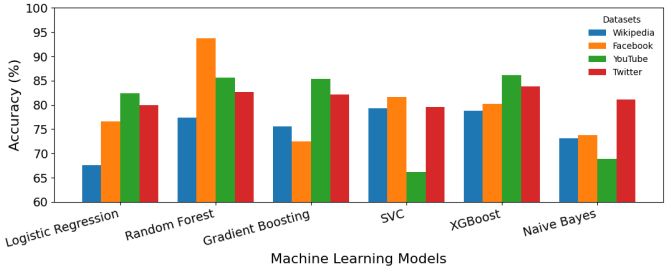

Fig. 2. Accuracy for the four datasets for machine learning models.

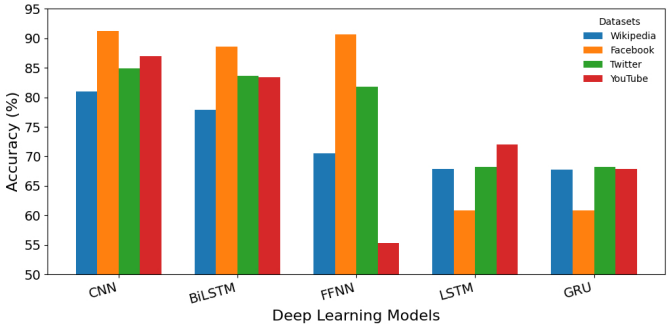

Fig. 3. Accuracy for the four datasets for deep learning models.

In the CNN + Random Forest (Model-Level Hybrid) approach, a CNN is used to generate semantic embeddings from embedded and padded text comments, which were then combined with traditional features, including TF-IDF, sentiment polarity, and comment length. The fused feature vector was passed to a Random Forest classifier. It was suitable for platforms like Wikipedia and Meta, where longer, structured content required robust analysis. Next was the CNN + XGBoost (Model-Level Hybrid) approach, where CNN-extracted embeddings were combined with XGBoost's efficient, regularized classification. This hybrid was ideal for high-throughput moderation tasks on platforms like YouTube and X, offering low-latency, high-accuracy detection of harmful content.

Finally, we tested CNN + Logistic Regression, again a model-level hybrid approach, in which CNN extracted high-level features from embedded and padded text, which were directly passed to a Logistic Regression classifier. We found it to be suitable for real-time settings, particularly on platforms like Meta.

D. Confidence-based Moderation Framework

To move beyond conventional binary moderation systems that block or allow content, we proposed a framework that incorporates a confidence-based mechanism. Instead of making rigid decisions based on raw output, the model's prediction confidence is used to inform graduated moderation actions. Each prediction is assigned to one of three confidence tiers: High Confidence (>75%): The model confidently identifies harmful content, triggering automatic blocking without delay. It applies to clearly toxic or abusive messages with minimal ambiguity. Moderate Confidence (55–75%): The content appears potentially harmful but falls into a gray area. Users receive warnings or educational notices, and the content is sent to a review queue where human moderators can make the final decision or handle appeals. Low Confidence (<55%): For content where the model lacks confidence in its judgment, the system allows it to be published by default, but it is flagged for passive monitoring or later review.

IV. Results

This study used four datasets sourced from publicly available Kaggle repositories, each representing a major social media platform. These datasets were already labeled ("pre-labeled") for binary classification. To evaluate the performance of the models, six key classification metrics were used: accuracy, precision, recall, F1-score, training time, and prediction time. These metrics collectively assessed each model's effectiveness in detecting cyberbullying while maintaining computational efficiency. Due to space constraints, only accuracy plots are presented. Accuracy represents the overall correctness of predictions, computed as the ratio of correct predictions to the total samples. Each dataset was split using an 80:20 ratio for training and testing. To handle class imbalance and prevent overfitting, SMOTE was applied during training along with dropout, early stopping, and class weighting for deep models.

Fig. 2 shows accuracy results for six classical ML classifiers: Logistic Regression, Random Forest, Gradient Boosting, Support Vector Classifier, XGBoost, and Naive Bayes. For Twitter (X), XGBoost achieved the highest accuracy (83.83%), followed by Random Forest (82.61%). For Facebook, Random Forest outperformed all models with a remarkable 93.70% accuracy and a high F1-score (0.92). For Wikipedia, Linear SVC (79.26%) and XGBoost (78.75%) led in accuracy. Logistic Regression lagged due to a weaker F1-score (0.65). For YouTube, XGBoost reached the highest accuracy (86.15%), but all models struggled with cyberbullying detection, highlighting limitations in precision and recall. These outcomes affirm that ensemble models like XGBoost and Random Forest offer a robust foundation for further deep and hybrid modeling.

Fig. 3 displays results for standalone deep learning architectures: FFNN, CNN, LSTM, GRU, and BiLSTM. For Twitter, CNN achieved the highest accuracy (84.87%), followed closely by BiLSTM (83.65%). FFNN also performed well (81.81%), while LSTM and GRU underperformed significantly (68.25%). For Facebook, CNN again led (91.30%), followed by FFNN (90.65%) and BiLSTM (88.55%). LSTM and GRU showed poor generalization ($\approx$ 60.88%). For Wikipedia, CNN maintained top performance (80.97%), followed by BiLSTM (77.84%). For YouTube, CNN achieved 87.01% accuracy, though its recall for cyberbullying was poor (0.04). BiLSTM followed with 83.41% accuracy but similarly low F1-score. These results highlight CNN's spatial feature extraction advantage, while BiLSTM's context modeling offered competitive results, justifying their use in hybrid architectures.

Fig. 4 presents results for the Hybrid Logistic Regression model, which achieved 96.08% accuracy on Wikipedia, 78.8% on Facebook, and 83.89% on Twitter. This model offered consistent performance with low training and prediction time, making it suitable for real-time moderation. Fig. 5 displays the

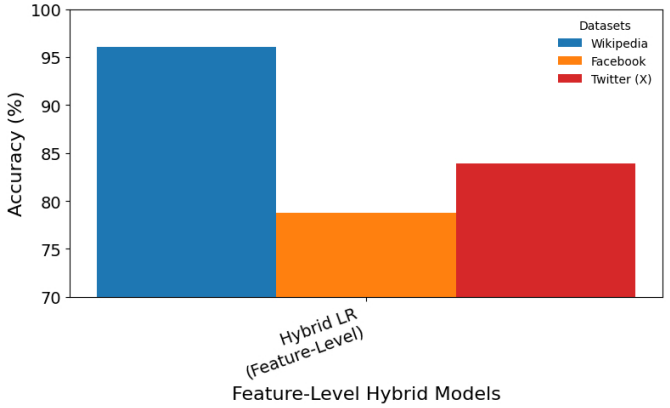

Fig. 4. Accuracy for the hybrid logistic regression approach.

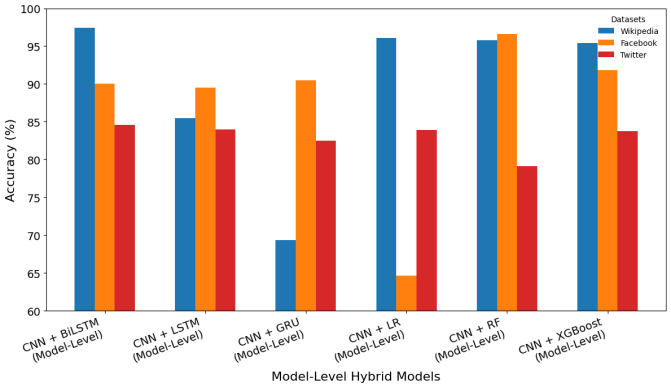

Fig. 5. Accuracy for the model level hybrid approaches.

results for model-level hybrids, where CNN embeddings were passed to secondary classifiers like BiLSTM, LSTM, GRU, Random Forest, XGBoost, and Logistic Regression. CNN + BiLSTM outperformed all models with 97.45% accuracy on Wikipedia, 90.03% on Facebook, 84.60% on Twitter.

Other models like CNN + LSTM and CNN + GRU performed well, but not as consistently. Notably, CNN + GRU achieved only 69.36% for the Wikipedia dataset, confirming its limited generalization in long-form text. All model-level hybrids outperformed their standalone deep learning counterparts. This shows the effectiveness of combining CNN's local pattern detection with temporal models or ensemble classifiers for nuanced detections.

Platform-Specific Observations: For the Wikipedia dataset, the best approaches were CNN + BiLSTM (97.45%) and Hybrid Logistic Regression (96.08%). For the Facebook dataset, Random Forest (93.70%) and CNN + BiLSTM (90.03%) stood out for longer, well-structured comments. CNN + RF also scored 96.58%. For the Twitter (X) dataset, peak hybrid accuracy was 84.60% (CNN + BiLSTM), but results remained slightly lower due to short, sarcastic, and informal text. Finally, for YouTube, we excluded this dataset from the hybrid evaluation due to extreme class imbalance and poor quality of comments. CNN (87.01%) was the best approach among all DL models but lacked reliability in detecting the minority class.

V. Conclusion

This study delivers a comprehensive framework for cyberbullying detection, rigorously evaluating Machine Learning, Deep Learning, and Hybrid Models across diverse platforms, including Wikipedia Talk, Twitter (X), Facebook (Meta), and YouTube datasets. A confidence-based mitigation module was introduced—translating prediction certainty into dynamic responses such as warnings, blocks, or content flags—marking a significant step toward real-world implementation. This work highlights the value of platform-specific evaluation, linguistic adaptability, and the ethical dimensions of automated moderation. Future extensions may explore multilingual and multimodal inputs, on-device deployment, and ethical safeguards to enhance system effectiveness and inclusivity further. Ultimately, the study supports advancing safer, more inclusive, and resilient digital communities.

References

[1] A. Alabdulwahab, M. A. Haq, and M. Alshehri, "Cyberbullying detection using machine learning and deep learning," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 10, 2023. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2023.0141045

[2] M. T. Hasan, M. A. E. Hossain, M. S. H. Mukta, A. Akter, M. Ahmed, and S. Islam, "A review on deep-learning-based cyberbullying detection," *Future Internet*, vol. 15, no. 5, 2023. [Online]. Available: https://www.mdpi.com/1999-5903/15/5/179

[3] J. O. Atoum, "Cyberbullying detection neural networks using sentiment analysis," in *2021 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2021, pp. 158–164.

[4] A. Perera and P. Fernando, "Cyberbullying detection system on social media using supervised machine learning," *Procedia Computer Science*, vol. 239, pp. 506–516, 2024, cENTERIS â International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050924014431

[5] N. Ejaz, F. Razi, and S. Choudhury, "Towards comprehensive cyberbullying detection: A dataset incorporating aggressive texts, repetition, peerness, and intent to harm," *Computers in Human Behavior*, vol. 153, p. 108123, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0747563223004740

[6] M. Priyadharshini, H. J, A. F. Banu, B.Nithya, S. K, and V.Murugesh, "Advanced cyberbullying detection: A hybrid model integrated with naïve bayes," in *2024 Third International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*, 2024, pp. 1–5.

[7] N. H. T. et al., "ProTect: a hybrid deep learning model for proactive detection of cyberbullying on social media."

[8] A. Akhter, U. K. Acharjee, M. A. Talukder, M. M. Islam, and M. A. Uddin, "A robust hybrid machine learning model for bengali cyber bullying detection in social media," *Natural Language Processing Journal*, vol. 4, p. 100027, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2949719123000249

[9] A. K. G and D. Uma, "Detection of cyberbullying using machine learning and deep learning algorithms," *2022 2nd Asian Conference on Innovation in Technology (ASIANCON)*, pp. 1–7, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252849883

[10] M. Mahat, "Detecting cyberbullying across multiple social media platforms using deep learning," in *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, 2021, pp. 299–301.

[11] A. S. Susmitha and P. Pujari, "Sentiment analysis of cyberbullying data in social media," *arXiv preprint arXiv:2411.05958*, 2024.

[12] M. A. Alabdali AM, "A novel approach toward cyberbullying with intelligent recommendations using deep learning based blockchain solution."