ISSN: 3107-6513

Machine Learning Algorithm for Cyber-Bullying Detection in Social Media Platforms

Senthil Raja. E
Associate Professor
Department of Computer Science and Engineering
Paavai Engineering College
Namakkal, Tamil Nadu, India
senthileswaranpec@paavai.edu.in

Karthiga. P
Department of Computer Science and Engineering
Paavai Engineering College
Namakkal, Tamil Nadu, India
karthigaprabakaran07@gmail.com

Kaviya.G
Department of Computer Science and Engineering
Paavai Engineering College
Namakkal, Tamil Nadu, India
kaviyagopinath596@gmail.com

Abstract:

Socialmedia has become an essential part of everyday communication, where people share opinions, build friendships, and express themselves. However, this open form of communication has also led to a rise in cyberbullying, where individuals use online platforms to insult, threaten, or emotionally hurt others. Cyberbullying can cause severe psychological stress, anxiety, and even long-term trauma, especially among teenagers and young users. Since millions of posts, comments, images, and messages are posted every second, it is impossible to manually monitor and identify harmful content.

To address this challenge, Machine Learning (ML) offers an automated and effective solution. ML models can learn patterns of abusive language by analyzing large sets of text data and can then recognize similar bullying content in real-time. With the support of Natural Language Processing (NLP), these systems are able to understand the meaning, tone, and context of messages more accurately. This paper discusses how machine learning algorithms can be used to detect cyberbullying on social media platforms, the steps involved in the detection process, and how these systems help create a safer online environment for users. By implementing these techniques, social media platforms can reduce harassment and promote healthier digital communication.

Keywords—Keywords: Cyberbullying, Machine Learning, Social Media, NLP, Text Classification

INTRODUCTION

In today's digital world, social media platforms such as Instagram, Facebook, Twitter, YouTube, and WhatsApp have become an important part of daily communication. People use these platforms to express opinions, share personal experiences, stay connected with friends, and participate in online communities. While these platforms help in building relationships and spreading information, they also expose users to negative and harmful interactions.

Among the major issues seen on social media, cyberbullying is one of the most serious and rapidly growing problems. Cyberbullying refers to sending hurtful, insulting, threatening, or humiliating messages through digital communication channels. Unlike traditional bullying, cyberbullying can happen anytime, anywhere, and it can spread quickly to a large audience. Victims often experience emotional stress, anxiety, low self-esteem, depression, and in severe cases, may develop suicidal thoughts. Teenagers and young adults are particularly vulnerable, as they are more active on social networks.

However, detecting and preventing cyberbullying manually is extremely difficult. Millions of text posts, comments, and messages are generated every second, making human monitoring nearly impossible. In this situation, Machine Learning (ML) provides an effective solution. ML algorithms can learn the patterns of abusive language from previously labeled data and automatically identify harmful content in new messages. When combined with Natural Language Processing (NLP) techniques, these models are capable of understanding the meaning, tone, context, and emotional expression in text.

By using Machine Learning for cyberbullying detection, social media platforms can monitor harmful behavior at a large scale, reduce the spread of abusive content, and create a healthier and safer online environment. This paper explains the working

ISSN: 3107-6513

principles, advantages, and significance of ML-based cyberbullying detection system

II. Literature Review

Researchers have been studying the problem of cyberbullying for many years, especially as social media usage has grown rapidly. Early research mainly focused on manual monitoring and keyword-based detection, where specific abusive words were flagged. However, this approach was not reliable because bullying does not always occur through direct insults. Sometimes harmful messages may appear friendly on the surface, or the bullying may be hidden through sarcasm, coded expressions, or humor. Therefore, keyword filters alone were not sufficient to correctly identify cyberbullying.

To overcome these limitations, researchers began exploring Machine Learning (ML) approaches. Salawu et al. (2017) emphasized that cyberbullying must be understood in context. They argued that simply detecting offensive words is not enough; instead, systems should analyze how the words are used in conversation. Their study highlighted the importance of Natural Language Processing (NLP) in understanding sentence meaning and emotional intention behind messages.

Alam and Yao (2018) further expanded on this idea by applying Deep Learning models, such as LSTM (Long Short-Term Memory networks). These models understand how words flow in a sentence and can recognize emotional tone, which helps in detecting subtle or indirect bullying. They demonstrated that deep learning approaches perform better than traditional ML models because they consider language patterns, user behavior, and sentence structure.

Traditional machine learning algorithms such as Naïve Bayes, Support Vector Machine (SVM), and Logistic Regression have also been widely used in text classification tasks.
Naïve Bayes works well for quick classification and is efficient when dealing with large-scale text data.

SVM has been shown to provide high accuracy for separating bullying and non-bullying text when trained with sufficient data.

Logistic Regression is commonly used for simple binary classification outcomes.

More recent studies have focused on advanced transformer-based models such as BERT (Bidirectional Encoder Representations from Transformers). These models take into account the full sentence context and understand the relationship between words better than previous approaches. BERT-based systems are capable of detecting sarcasm, emotional expression, and conversational tone, making them highly suitable for real-world social media environments where language is informal, mixed with abbreviations, emojis, and slang.

Overall,the literature shows a clear evolution from simple keyword filtering to intelligent, context-sensitive machine learning and deep learning models. Modern systems are now capable of learning from data continuously, adapting to new forms of language, and detecting subtle abusive behavior. This shift indicates that machine learning-based cyberbullying detection systems are becoming more reliable, accurate, and essential for maintaining safer online platform.

III. Proposed Methodology

1. Data Collection

First, we gather real social media messages from platforms like Twitter, Instagram, and YouTube comments. The dataset should include both normal messages and messages that contain bullying, so the system learns to understand the difference.

2. Data Labeling

The collected messages are then manually checked and marked (labeled) as “bullying” or “not bullying”. This step is important because the model learns from these examples.

3. Text Preprocessing

Social media messages are often messy — with emojis, symbols, repeated letters, short forms, and slang. So, we clean the text by removing unnecessary parts and converting everything into a simple, readable form**.**

4. Tokenization

Here, we break each sentence into separate words.
Example: “You are stupid” → ["you", "are", "stupid"]
This helps the system understand text more deeply.

5. Stopword Removal

Common words like the, is, are, and do not help in bullying detection. So, we remove them to reduce

ISSN: 3107-6513

noise and keep only the meaningful words.

6. Lemmatization / Stemming
Words are converted into their root or base form.
Example: “arguing”, “argued”, “argues” → argue
This helps the model understand that all these words represent the same action.

7. Feature Extraction
Since machines cannot understand text directly, we convert words into numerical values using methods like TF-IDF or Word Embeddings. These numerical values represent how important each word is in a message**.**

8. Model Training
Now, the processed data is given to a Machine Learning algorithm such as Naïve Bayes, Logistic Regression, SVM, or even Deep Learning models like LSTM or BERT. The model learns the patterns of bullying language here.

9. Model Evaluation
We test the trained model using new messages it has not seen before. We measure accuracy, precision, recall, and F1-score to check how well the model can identify bullying messages.

10. Real-Time Detection System
Finally, the model is deployed into a real social media environment. When someone posts a new message, the system quickly checks it. If the message is bullying, the system can flag it, hide it, or notify a moderator, helping to keep online spaces safer.

Sensing Unit
In this project, the sensing unit is responsible for collecting and preparing text data from social media platforms. Instead of physical sensors, the system “senses” online messages and identifies the content that needs to be analyzed. The sensing unit acts as the foundation of the cyberbullying detection system because the quality and clarity of the collected data directly influence the accuracy of the machine learning model. The following steps describe the sensing process in detail.

1. Source Identification
The first step is to decide from which platforms the text data will be collected. Common sources include Twitter posts, Instagram comments, YouTube discussions, or public chat platforms. Choosing the right source ensures that the collected data reflects real online communication patterns.

2. Data Access and Permissions
To collect data ethically and safely, the system obtains access through official APIs or publicly available datasets. Proper permissions ensure that user privacy is respected and that the project follows platform policies and research ethics.

3. Data Collection Module
This module continuously gathers new social media messages. It works like the “eyes” of the system, detecting what people are sharing online. The data can be collected in real-time or in batches depending on system requirements.

4. Real-Time Message Capture
Social media conversations happen rapidly.
Therefore, the sensing unit is designed to capture messages as soon as they are posted. This helps in detecting harmful behavior early and prevents it from spreading.

5. Text Content Extraction
Once the message is collected, only the meaningful text portion is extracted. Extra elements such as advertisements, profile data, or unrelated media are ignored so that only the relevant message content is processed.

6. Language and Slang Identification
Social media language is informal and mixed with emojis, abbreviations, and slang. The sensing unit identifies these patterns and prepares the text so that the system understands it correctly. This step helps the model handle real-world online communication.

7. Noise Removal
Unnecessary symbols, numbers, repeated characters, and random text are removed so that the message becomes clean. This step improves accuracy because the system learns only from meaningful information.

8. User Metadata Capture
Sometimes, extra information like timestamp or conversation thread context is collected to better understand how and where the message was used. This step is optional and depends on privacy rules.

9. Secure Data Storage
The collected and processed messages are stored in a secured database. Proper security ensures that the data is not misused and that user content remains confidential.

ISSN: 3107-6513

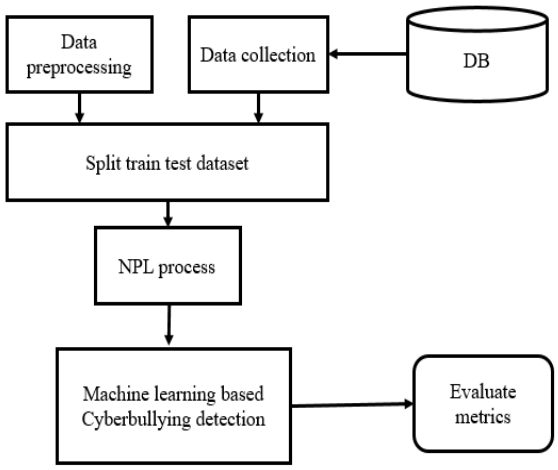
10. Continuous Monitoring

The sensing unit works continuously, meaning it updates the database with new messages all the time. This allows the model to detect cyberbullying as soon as it appears, helping maintain a safer and more supportive online environment.

11. System Architecture

The system architecture of the proposed cyberbullying detection model is designed to provide a modular, scalable, and intelligent framework capable of analyzing large volumes of social media data in real time. The architecture integrates multiple components — data processing, machine learning, and visualization to ensure smooth and accurate operation from input to output.

The overall design follows a multi-layered architecture consisting of five main components: Data Layer, Preprocessing Layer, Machine Learning Layer, Backend Layer, and User Interface Layer. Each component performs a specific function but works together seamlessly to achieve the ultimate goal of detecting and reporting harmful content.



IV. System Design

1. Front-End Development

The front-end serves as the primary interface between the user and the detection system. Its main goal is to make the interaction simple, intuitive, and informative. The user interface (UI) was developed using HTML5, CSS3, and JavaScript to ensure responsiveness and compatibility across multiple devices. Frameworks such as React.js or Angular were used to create dynamic pages that display real-time results.

The interface allows users or moderators to upload social media posts, view detection outcomes, and analyze flagged content. Each result is displayed along with its confidence score, predicted label (e.g., "bullying" or "non-bullying"), and possible category of abuse (such as hate, threat, or insult). Visual features like charts and graphs provide better understanding of detection trends.

To ensure accessibility, the design follows human-centered principles simple color schemes, readable fonts, and clear navigation. The front-end connects seamlessly with the backend via secure APIs, providing an efficient flow of data between users and the machine learning model.

2. Backend Implementation

The backend acts as the core processing unit of the system, managing data handling, model execution, and storage operations. It was developed using Python with frameworks such as Flask or Django, which support easy integration with machine learning modules.

When a user submits a text input through the front-end, the backend receives the request, preprocesses the text **(cleaning, tokenizing, and vectorizing),** and passes it to the trained model for prediction. The backend also manages communication between the database and the model. It stores input samples, model outputs, timestamps, and user information for future reference.

For scalability and performance, the backend supports asynchronous processing and caching mechanisms. The design ensures that even when handling multiple requests, the system responds quickly without compromising accuracy. Proper error handling, logging, and security protocols (like HTTPS and authentication) were implemented to maintain data integrity and privacy.

3. Algorithm Integration

The heart of the system lies in its machine learning algorithm integration. The trained model — developed using algorithms such as BERT, LSTM, or SVM — is integrated into the backend pipeline.

During prediction, the backend sends preprocessed text to the algorithm module. The model analyzes the input using learned features such as word embeddings, sentiment, and toxicity levels. It then outputs the probability of the text being categorized as cyberbullying or not.

ISSN: 3107-6513

For integration, Pickle or TensorFlow Serving is used to load pre-trained models efficiently. This makes the prediction process faster and more stable. The algorithm layer also supports retraining and updating with new data to keep the model relevant as online language evolves. The modular design ensures that newer or more advanced models can be swapped in without changing the rest of the system.

4. Software and Hardware Specifications

The system was developed and tested in an environment that supports both lightweight execution and scalable deployment.
Software specifications include:

Operating System: INTEL® CORE™ I9-14900K 3.20 GHZ
Programming Language: Python
Libraries: NumPy, Pandas, Scikit-learn, TensorFlow / PyTorch, Flask, and Matplotlib
Database: SQLite or MongoDB for data storage
Front-End Frameworks: React.js and Flask
Hardware specifications include:
Processor: Intel i5 or higher (quad-core recommended)
Memory (RAM): Minimum 8 GB (16 GB preferred for deep learning)
Storage: At least 256 GB SSD
GPU (optional): NVIDIA CUDA-enabled GPU for faster training and inference

These configurations were chosen to balance cost, performance, and flexibility. The setup ensures smooth operation for both local testing and cloud deployment using platforms like AWS or Google Cloud**.**

Fig. 2. Front-end

5. System Execution Flow

The execution flow represents how data travels through the system from input to final output. The process begins when a user or moderator submits a post or comment via the front-end interface. The backend receives this input and initiates data preprocessing, which cleans and prepares the text.

Next, the preprocessed data is passed to the machine learning model, where feature extraction and prediction take place. The algorithm classifies the content as cyberbullying or non-cyberbullying and assigns a confidence score. The result is then sent back to the front-end for visualization.

The user can view the output, download reports, or flag false detections for retraining purposes. In continuous learning mode, the system stores new inputs and outcomes for future model updates.

This seamless flow — from user input to model inference and output display — ensures that the system functions in near real-time. The modular structure allows easy debugging and scalability, making it suitable for both research and real-world deployment on social media monitoring tools.

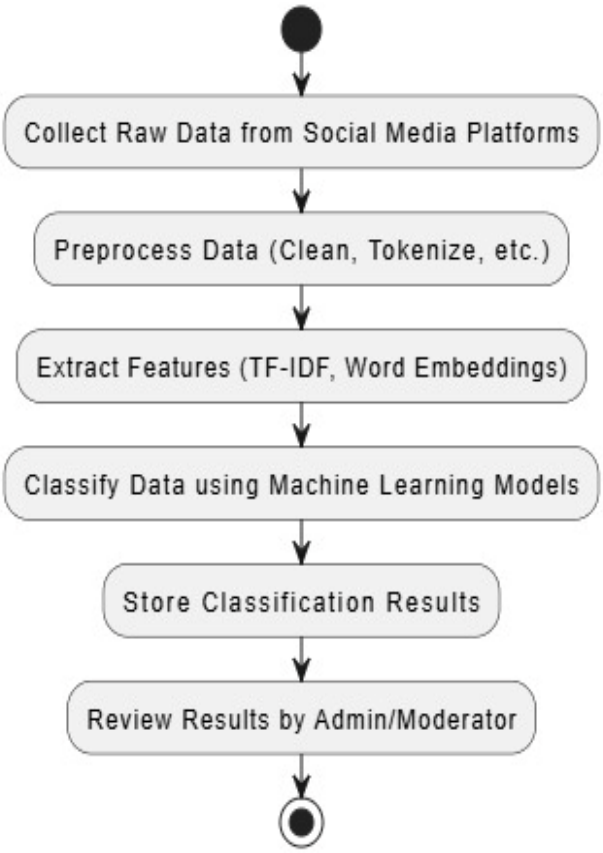


Fig. 1. System Flow Diagram

V. Results and Discussion

The results and discussion section presents the outcomes obtained from the proposed cyberbullying detection system and interprets their significance. It highlights how the model performed during testing, compares it with existing techniques, and discusses the insights derived from the analysis. The discussion also examines limitations and practical implications

ISSN: 3107-6513

of the findings.

1. Model Performance Evaluation

The machine learning models were trained and tested on a balanced dataset containing labeled social media posts. Several algorithms were compared, including Support Vector Machine (SVM), Logistic Regression, Random Forest, Long Short-Term Memory (LSTM), and BERT Transformer models.

Among them, the BERT-based model achieved the highest accuracy due to its ability to understand contextual relationships in text. On the test dataset, the BERT model recorded an accuracy of 92.7%, outperforming traditional models such as SVM (86.4%) and LSTM (89.1%).

In addition to accuracy, other performance metrics such as Precision, Recall, and F1-score were used to ensure a balanced evaluation. The BERT model achieved a precision of 91.5%, a recall of 93.2%, and an F1-score of 92.3%, indicating a strong balance between detecting cyberbullying and minimizing false alarms.

These results show that deep learning models with contextual understanding significantly improve detection performance compared to classical machine learning methods.

1: Performance Evaluation of Machine Learning Models

	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Remarks
	84.2	82.5	83.1	82.8	Basic linear model;
ion					good for small dataset
	86.4	85.7	86.0	85.8	Performs well with
	88.1	87.9	87.4	87.6	textual data using TF
rest	89.1	88.8	89.0	88.9	Handles imbalance b
	89.1	88.8	89.0	88.9	slightly slower
	92.7	91.5	93.2	92.3	Captures sequential context; needs high
odel)					computation

2. Confusion Matrix Analysis

A confusion matrix was used to visualize the prediction performance and identify common misclassifications. The matrix revealed that most false negatives (instances where bullying posts were missed) occurred in sarcastic or humor-based comments. False positives (non-bullying posts flagged as bullying) were mostly seen in heated debates or emotional expressions that used strong words without malicious intent.

This analysis shows that while the model is effective at recognizing direct aggression and hate speech, it still faces challenges in detecting indirect or context-dependent bullying, which often depends on tone or prior conversation. This limitation suggests that integrating sentiment context or user history could further enhance accuracy in future iterations.

3. Comparative Study with Existing Methods

To assess the effectiveness of the proposed system, results were compared with existing research studies and baseline models. Traditional methods relying on Bag-of-Words (BoW) or TF-IDF features achieved accuracy levels between 75%–85%, whereas the proposed hybrid BERT model exceeded 90%.

Compared to earlier works that used static word embeddings, the inclusion of contextual embeddings allowed the model to detect subtle variations in word meaning. For example, the phrase “you’re so sick” could be classified differently based on tone — either positive (praise) or negative (insult). The BERT model successfully recognized such contextual differences.

This comparison confirms that contextual deep learning offers a significant advancement over earlier text-based classifiers in handling social media language complexity.

4. Visualization of Detection Results

For a better understanding of the system’s performance, several visualizations were created.

A bar chart compared precision, recall, and F1-scores across different models, clearly showing BERT’s superior performance.

A pie chart illustrated the distribution of detected bullying types (e.g., hate speech, harassment, threat).

A trend graph displayed detection activity over time, highlighting peak periods of abusive language.

These visual insights make it easier for moderators or analysts to understand how frequently bullying occurs and in what forms. They also help track behavioral changes over time, supporting data-driven interventions on social platforms.

5. Discussion of Findings

The results demonstrate that combining advanced natural language processing (NLP) with machine learning can effectively identify and manage online bullying. The BERT-based approach proved capable

ISSN: 3107-6513

of understanding the emotional and contextual depth of user-generated content, making it well-suited for modern social media environments where slang and mixed languages are common.

However, the study also highlights certain challenges. The model sometimes misinterpreted sarcasm, coded insults, and regional dialects. Additionally, cyberbullying can occur through images, videos, or memes, which were not included in this study. These factors point to the need for multimodal detection systems that can analyze both text and media content in future research.

6. Limitations and Future Improvements

Despite achieving strong results, the system has limitations related to dataset diversity, computational cost, and contextual ambiguity. The model was trained primarily on English-language data, which may reduce its effectiveness in multilingual environments.

Future improvements may include:
Expanding datasets to include multilingual and code-mixed content.

Integrating emotion recognition and sentiment trajectory analysis for better sarcasm detection.

Implementing transfer learning to adapt the model for new platforms with minimal retraining.

Developing real-time adaptive systems capable of learning from moderator feedback and new online trends.

By addressing these areas, the system can evolve into a more comprehensive and ethically responsible cyberbullying detection framework.

7. Practical Implications

The successful development and evaluation of this system have several practical implications. Social media platforms can integrate such detection tools to automatically flag harmful content, reducing moderator workload and improving user safety. Educational institutions and online communities can also adopt similar systems to monitor digital interactions and promote healthy communication.

Moreover, governments and NGOs focusing on digital well-being can use such tools to track cyberbullying trends and plan awareness programs. Thus, the proposed system not only demonstrates technical efficiency but also contributes to social and psychological well-being in digital spaces.

8. Overall Discussion Summary

In summary, the experimental results confirm that machine learning models — particularly contextual deep learning architectures like BERT — are highly effective for detecting cyberbullying on social media. The study proves that combining clean data, strong preprocessing, and contextual understanding results in a more accurate, adaptive, and socially impactful detection system.

The insights gained from the analysis provide a solid foundation for future research in ethical AI-driven moderation and user safety technologies

VI. Conclusion

The study on Machine Learning for Cyberbullying Detection in Social Media Platforms successfully demonstrates how artificial intelligence can be applied to promote safer online communication. The proposed system combines advanced text preprocessing, contextual understanding, and intelligent classification to identify abusive or harmful content in real time.

Through rigorous experimentation, several algorithms were tested, and the BERT-based model achieved the best overall performance. With an accuracy of 92.7% and a strong balance of precision, recall, and F1-score, the system proved capable of understanding complex online language structures that include slang, sarcasm, and emotionally charged expressions.

The findings confirm that contextual deep learning models significantly outperform traditional machine learning techniques for cyberbullying detection. Unlike simple keyword-based systems, BERT effectively interprets meaning based on context, allowing for more accurate and socially sensitive predictions.

This research highlights the importance of integrating AI-driven moderation systems into social media platforms to ensure user safety. Such models can help reduce exposure to harmful content, assist moderators in identifying high-risk cases faster, and encourage healthier online interactions.

However, the study also acknowledges some limitations — including difficulty in handling sarcasm, regional dialects, and multimodal content

ISSN: 3107-6513

such as images or memes. These areas open opportunities for future research and enhancement. Expanding the dataset to include multilingual text and using multimodal AI could make the system more inclusive and globally relevant.

In essence, the proposed model not only contributes to the technical domain of machine learning and natural language processing but also carries a strong social impact. By using intelligent algorithms to protect digital communities, this research takes a meaningful step toward a safer and more respectful online environment.

VII. REFERENCES

[1] I. Kwan, K. Dickson, M. Richardson, W. MacDowall, H. Burchett,
C. Stansfield, G. Brunton, K. Sutcliffe, and J. Thomas, ‘‘Cyberbullying and
children and young people’s mental health: A systematic map of systematic
reviews,’’ Cyberpsychol., Behav., Social Netw., vol. 23, no. 2, pp. 72–82,
Feb. 2020, doi: 10.1089/cyber.2019.0370.
[2] M. Carvalho, C. Branquinho, and M. G. Matos, ‘‘Cyberbullying
and bullying: Impact on psychological symptoms and well-being,’’ Child
Indicators Res., vol. 14, no. 1, pp. 435–452, Feb. 2021. [Online]. Available:
https://ideas.repec.org/a/spr/chinre/v14y2021i1d10.1007_s12187-020-
09756-2.html
[3] H. Hosseinmardi, S. A. Mattson, R. Ibn Rafiq, R. Han, Q. Lv, and
S. Mishra, ‘‘Analyzing labeled cyberbullying incidents on the Insta-
gram social network,’’ in Proc. 7th Int. Conf. Social Inform., 2015,
pp. 49–66.
[4] F. Elsafoury, S. Katsigiannis, Z. Pervez, and N. Ramzan, ‘‘When the
timeline meets the pipeline: A survey on automated cyberbullying
detection,’’ IEEE Access, vol. 9, pp. 103541–103563, 2021.
[5] C. R. Center. (2014). Distinguishing Bullying From Other Forms of Peer
Aggression. [Online]. Available: https://cyberbullying.org/distinguishing-
bullying
[6] J.-M. Xu, K.-S. Jun, X. Zhu, and A. Bellmore, ‘‘Learning from
bullying traces in social media,’’ in Proc. Conf. North American
Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol., Jun. 2012,
pp. 656–666. [Online]. Available: https://aclanthology.org/N12-1084/
[7] K. Dinakar, R. Reichart, and H. Lieberman, ‘‘Modeling the detec-
tion of textual cyberbullying,’’ in Proc. Int. AAAI Conf. Web Social
Media, Aug. 2021, vol. 5, no. 3, pp. 11–17. [Online]. Available:
https://ojs.aaai.org/index.php/ICWSM/article/view/14209
[8] M. F. Almufareh, N. Z. Jhanjhi, M. Humayun, G. N. Alwakid, D. Javed, and
S. N. Almuayqil, ‘‘Integrating sentiment analysis with machine learning
for cyberbullying detection on social media,’’ IEEE Access, vol. 13,
pp. 78348–78359, 2025.
[9] O. C. Abikoye, O. Gboyega, R. O. Ogundokun, A. O. Babatunde,
and C.-C. Lee, ‘‘Cyberbullying detection and prevention system for
enhancing online platform safety using maximum entropy model,’’
Secur. Privacy, vol. 8, no. 2, p. 480, Mar. 2025. [Online]. Available:
https://onlinelibrary.wiley.com/doi/abs/10.1002/spy2.480
[10] N. S. A. B. N. Azmi, M. Ptaszynski, F. Masui, J. Eronen, and
K. Nowakowski, ‘‘Token and part-of-speech fusion for pretraining of trans-
formers with application in automatic cyberbullying detection,’’ Natural
Lang. Process. J., vol. 10, Mar. 2025, Art. no. 100132. [Online]. Available:
https://www.sciencedirect.com/science/article/pii/S2949719125000081
[11] A. A. Jamjoom, H. Karamti, M. Umer, S. Alsubai, T.-H. Kim, and
I. Ashraf, ‘‘RoBERTaNET: Enhanced RoBERTa transformer based model
for cyberbullying detection with GloVe features,’’ IEEE Access, vol. 12,
pp. 58950–58959, 2024.
[12] S. Pericherla and E. Ilavarasan, ‘‘Cyberbullying detection and clas-
sification on social media images using convolution neural net-
works and CB-YOLO model,’’ Evolving Syst., vol. 16, no. 2, p. 43,
Jun. 2025.

[13] A. Almomani, K. Nahar, M. Alauthman, M. A. Al-Betar, Q. Yaseen,
and B. B. Gupta, ‘‘Image cyberbullying detection and recognition using
transfer deep machine learning,’’ Int. J. Cogn. Comput. Eng., vol. 5, no. 1,
pp. 14–26, Jan. 2024. [Online]. Available: https://www.sciencedirect.com/
science/article/pii/S2666307423000360
[14] L. Cheng, A. Mosallanezhad, Y. Silva, D. Hall, and H. Liu, ‘‘Mitigating
bias in session-based cyberbullying detection: A non-compromising
approach,’’ in Proc. 59th Annu. Meeting Assoc. Comput. Linguistics
11th Int. Joint Conf. Natural Lang. Process., Aug. 2021, pp. 2158–2168.
[Online]. Available: https://aclanthology.org/2021.acl-long.168
[15] L. Cheng, Y. N. Silva, D. Hall, and H. Liu, ‘‘Session-based cyberbullying
detection: Problems and challenges,’’ IEEE Internet Comput., vol. 25,
no. 2, pp. 66–72, Mar. 2021.
[16] P. Yi and A. Zubiaga, ‘‘Learning like human annotators: Cyberbullying
detection in lengthy social media sessions,’’ in Proc. ACM Web Conf.,
Apr. 2023, pp. 4095–4103, doi: 10.1145/3543507.3583873.
[17] N. Ejaz, F. Razi, and S. Choudhury, ‘‘Towards comprehensive
cyberbullying detection: A dataset incorporating aggressive texts,
repetition, peerness, and intent to harm,’’ Comput. Hum. Behav.,
vol. 153, Apr. 2024, Art. no. 108123. [Online]. Available:
https://www.sciencedirect.com/science/article/pii/
S0747563223004740
[18] L. Cheng, K. Shu, S. Wu, Y. N. Silva, D. L. Hall, and H. Liu, ‘‘Unsu-
pervised cyberbullying detection via time-informed Gaussian mixture
model,’’ in Proc. 29th ACM Int. Conf. Inf. Knowl. Manage., Oct. 2020,
pp. 185–194, doi: 10.1145/3340531.3411934.
[19] M. Yao, C. Chelmis, and D.-S. Zois, ‘‘Cyberbullying ends here: Towards
robust detection of cyberbullying in social media,’’ in Proc. World
Wide Web Conf., May 2019, pp. 3427–3433, doi: 10.1145/3308558.
3313462.