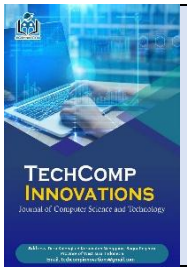

Journal of Computer Science and Technology, Vol 2 (1) 2025: 1-15

TechComp Innovations

https://mabadiiqtishada.org/index.php/TechCompInnovations

ISSN: 3062-8903; e-ISSN: 3062-889X

AI-Powered Frameworks for the Detection and Prevention of Cyberbullying Across Social Media Ecosystems

Md. Anas Mondol[1], Md. Ashaf Uddaula[2], Md. Safaet Hossain[3*], Mst. Ayesha Siddika[4]

[1] *Department of Computer Science and Engineering. City University. Bangladesh*

[2] *Department of Computer Science and Engineering. Daffodil International University. Bangladesh*

[3] *Department of Computer Science & Engineering. University of Information Technology and Sciences. Bangladesh*

[4] *Department of Computer Science and Engineering. Prime University. Dhaka. Bangladesh*

ARTICLE INFO

Received May 27, 2025
Accepted June 14, 2025
Published June 18, 2025

Keywords:
Cyberbullying Detection; Text Classification; Machine Learning; Deep Learning; Social Media

ABSTRACT

This study presents an advanced AI-powered framework to detect and prevent cyberbullying across diverse social media platforms using a multiclass classification approach. Addressing the growing complexity and linguistic diversity of online abuse, the research integrates various machine learning (RF, LR, SVM) and deep learning (Bi-LSTM, BERT) models trained on a balanced dataset covering bullying categories based on religion, age, ethnicity, gender, and neutral content. Data preprocessing, tokenization, feature extraction via TF-IDF and CountVectorizer, and class balancing using SMOTE were applied to enhance model accuracy. The proposed system further supports real-time detection through social media APIs, offering dynamic monitoring and intervention capabilities. Among the tested models, Random Forest and BERT achieved the highest classification performance with 94% accuracy. Despite its robust architecture, limitations include dependence on English-language datasets, exclusion of multimodal data (e.g., memes, audio), and API restrictions that challenge scalability. Future development will focus on incorporating vision-language models and optimizing the system for real-time, multilingual, and multimodal environments. This study contributes to digital safety efforts by proposing a scalable and adaptive detection system suitable for safeguarding users from evolving forms of online harassment.

How to cite: Mondol, M. A., Uddaula, M. A., Hossain, M. S., & Siddika, M. A. (2025). AI-Powered Frameworks for the Detection and Prevention of Cyberbullying Across Social Media Ecosystems. *TechComp Innovations: Journal of Computer Science and Technology*, 2(1), 1–15. https://doi.org/10.70063/techcompinnovations.v2i1.59

This is an open-access article distributed under the Creative Commons 4.0 Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. ©2023 by author.

* *Corresponding author: safayeth@gmail.com*

INTRODUCTION

In the age of hyperconnectivity, social media has revolutionized human communication and interaction, enabling real-time exchanges across geographic and cultural boundaries. However, the increased use of these platforms has also given rise to new forms of harm, one of the most prevalent being cyberbullying. This digital abuse is characterized by persistent harassment, offensive language, and targeted attacks, often based on gender, religion, ethnicity, or age, resulting in severe psychological consequences such as anxiety, depression, and in extreme cases, suicide (**Kowalski et al., 2021**). The global COVID-19 pandemic further intensified these issues, as prolonged screen time and limited in-person interactions created a breeding ground for online hostility (**Giumetti et al., 2022**).

While numerous studies have proposed machine learning (ML) and deep learning (DL) solutions for cyberbullying detection, many are constrained by narrow linguistic, cultural, or platform-specific scopes. For example, existing approaches often rely on datasets limited to a particular language or region, such as Bengali, Arabic, or Moroccan dialects, which restricts their generalizability (**Rachidi et al., 2023; Beniwal et al., 2023**). Furthermore, models designed for one platform, such as Instagram or Twitter, may not perform well across other platforms with differing content structures and user behaviors. As a result, current frameworks often lack the scalability and flexibility required to address the dynamic and multi-platform nature of cyberbullying in real-world contexts (**Sharma et al., 2020**).

This study addresses these critical limitations by proposing an AI-powered framework capable of detecting and preventing cyberbullying across diverse social media ecosystems using multiclass classification algorithms. The framework integrates both ML models—such as Random Forest (RF), Logistic Regression (LR), and Support Vector Machines (SVM)—and DL models like Bidirectional Long Short-Term Memory (Bi-LSTM) and Bidirectional Encoder Representations from Transformers (BERT). These models are trained and validated on a balanced dataset derived from publicly available sources, categorized into multiple dimensions of cyberbullying: religion, age, gender, ethnicity, and neutral content (**Mondol et al., 2025**). The framework also implements real-time content analysis through social media APIs, enhancing its relevance and responsiveness in dynamic online environments.

The urgency of developing an adaptable and high-performing cyberbullying detection system lies in the growing scale and sophistication of online abuse. Recent reports indicate a substantial increase in cyberbullying cases globally, with adolescents and marginalized communities being disproportionately affected (**Almeida et al., 2023**). Moreover, manual moderation by platforms is often insufficient due to the sheer volume of content and the subtleties of language that may obscure harmful intent (**Soni & Singh, 2022**). Thus, there is a pressing need for intelligent systems that not only detect cyberbullying with high accuracy but also provide scalable and language-independent solutions. This research aims to fill a critical gap by introducing a unified, AI-driven framework that combines technical rigor with practical adaptability. By leveraging advanced classification models and real-time data processing capabilities, this study contributes to the ongoing effort to make digital spaces safer and more inclusive for all users.

RESEARCH METHOD

This study focuses on building an ML model capable of identifying cyberbullying material across various types of real-time social media content. The dataset utilized for this purpose is a publicly available Kaggle dataset (**Kaggle, 2022**), which provides labeled data for training ML models. Dataset Description: The dataset consists of two primary dimensions: (1) Text

Data: This includes raw text content from social media platforms and (2) Cyberbullying Categories: The labels are organized into five categories: (a) Religion: Content targeting religious beliefs or communities; (b) Age: Posts involving discriminatory or harmful language based on age; (c) Ethnicity: Text directed toward ethnic groups or minorities; (d) Gender: Content with harmful language or biases related to gender; and (e) Not Cyberbullying: Neutral content that does not contain any abusive or harmful language.

Table I: Examples of Text and Corresponding Cyberbullying Categories

Text	Category
"Women can't drive properly!"	Gender
"Go back to your country!"	Ethnicity
"Religion is just a scam!"	Religion
"You're too old to get this meme!"	Age
"What a beautiful day it is today!"	Not bullying

During the phase of the data processing known as preprocessing, the NLTK is used. The NLTK tool is used for a variety of text processing tasks, including tokenization of text patterns, removal of stop words from text, etc., (**Pericherla, et.al., 2020**)

a. Lowering Text: It lowercase each letter in each word in the tokenization list. Before lowering "Hello" and after lowering "hello" (**Patidar, 2021**)
b. Removing Stop Words: This is the most crucial aspect of preprocessing. Stop terms are meaningless data phrases. Using NLTK, stop words can be eliminated very effortlessly. In this stage, stop words such as \t, https, and \u are eliminated from the text (**Romsaiyud, et.al., 2017**)
c. Tokenization: Tokenization involves dividing an input text into smaller pieces, with the resulting words and phrases being added to a separate list (**Thangarasu, et.al., 2023**). We begin by converting the tokenized text into sentences using PunktSentenceTokenizer (**Gupta, et.al., 2023**). The phrases are then tokenized using one of four tokenizers, which separate the sentences into individual words: Whitespace Tokenizer, Word Tokenizer, PunctWord Tokenizer, WordPunct Tokenizer, and Sentence Tokenizer.
d. WordNet Lemmatizer: The WordNet Lemmatizer is able to determine a word's synonyms, as well as its meaning and a number of other related concepts, and then connect all of these concepts to the original word (**Zhang, et.al., 2019**).

Feature Extraction

In this phase, the proposed model transforms the data into an appropriate format before passing it to the ML models (**Alothali, et.al., 2021**). The features of the data are encoded and stored within the dataset. The TFIDF and Count Vectorizer are utilized to derive the data's features (**Balet, et.al., 2023**). After this phase has been completed, SMOTE will rebalance the training set because unbalanced data may bias the training of the model (**Pradhan, et.al., 2020**).

Proposed Model

This research proposes a way to identify cyberbullying on social media. The primary distinction with earlier research is that, in addition to creating an ML model to identify cyberbullying content, it also put that model into practice in specific areas using real-time

content via social media APIs such as Twitter, TikTok, and Instagram Graph APIs. The whole process to identify and stop cyberbullying on social media can be broken down into the first two stages: setting up an experiment and creating a model.

1. Setting up an Experiment

Methods of identifying cyberbullying include a sequential process including RF, LR, DT, SVM, Bi-LSTM, and BERT. The following steps detail the experimental workflow:

a. Dataset Collection: Currently, we use open-source Kaggle data to obtain a certain amount of social media information from a chosen category (**Kaggle, 2023**).
b. Data Preprocessing: It employs state-of-the-art data preprocessing, data analysis, extraction methods, and SMOTE to enhance the retrieved data and ensure its highest quality and most applicable results.
c. Feature Extraction: The preprocessed text is sent to sophisticated ML models (see the section on Creating a Model) to determine the probability of the social media data obtained, deciding the category in which the text should be placed.
d. Probabilistic Evaluation: Probability serves as our primary evaluation metric, addressing the multiclass character of cyberbullying classification in an elegant manner. It achieves a balance between accuracy and memorization, accurately depicting the model's performance nuances (**Matias, et.al., 2023; Pawar, et.al., 2023**).
e. Contextual Refinement: This method reveals a new layer of context whenever a piece of material receives a probability indicating it may be used in cyberbullying. It routinely extracts the user's most recent five pieces of social media material from their timeline. These materials help us better comprehend the user's online activities, thereby improving the reliability of our evaluations.
f. Threshold-based Decision: The average probability of the user's posts is calculated. If this probability exceeds a predetermined threshold (e.g., 0.5), the content is flagged as potentially harassing and recorded in a database. Conversely, content with probabilities below the threshold is not flagged, and any prior records are removed. This database-driven system facilitates the systematic tracking of flagged content for potential intervention.
g. Time Cost Considerations: While the sequential application of multiple ML methods (RF, LR, DT, SVM, Bi-LSTM, BERT) and feature extraction techniques like TF-IDF or Count Vectorization adds computational overhead, the system is designed to optimize time costs by dynamically selecting the best-performing model and feature combination through comparative analysis. This ensures efficiency without compromising accuracy. Future improvements may incorporate lightweight models or ensemble approaches to further minimize time costs.

By using probability as a guide and developing a database-driven system, it brings a cutting-edge approach to the problem of identifying cyberbullying. This combination of state-of-the-art methods and original database integration illustrates the dedication to both efficiency and novelty.

In Fig. 1, the two main components of the architecture diagram for social media cyberbullying detection are NLP and ML. In real-time, posts, comments, and opinions about cyberbullying on social media are gathered in the English language for the study (**Singh, et.al., 2023**). The use of data preprocessing techniques gets the data ready for ML algorithms. Features like TF-IDF and count vectorizer are applied for the feature selection of text after the dataset is divided into training and testing sets (**Balet, et.al., 2023**). In order to prevent bias in

the model's training, SMOTE rebalances the training set after this phase is complete (Pradhan, et.al., 2020).

The accuracy of each model was assessed by referencing performance metrics and evaluation criteria reported in prior studies, as identified in a comprehensive literature survey. This included a comparison of published results for RF, LR, DT, SVM, Bi-LSTM, and BERT algorithms, focusing on metrics such as accuracy, precision, recall, and F1 scores within similar cyberbullying detection contexts.

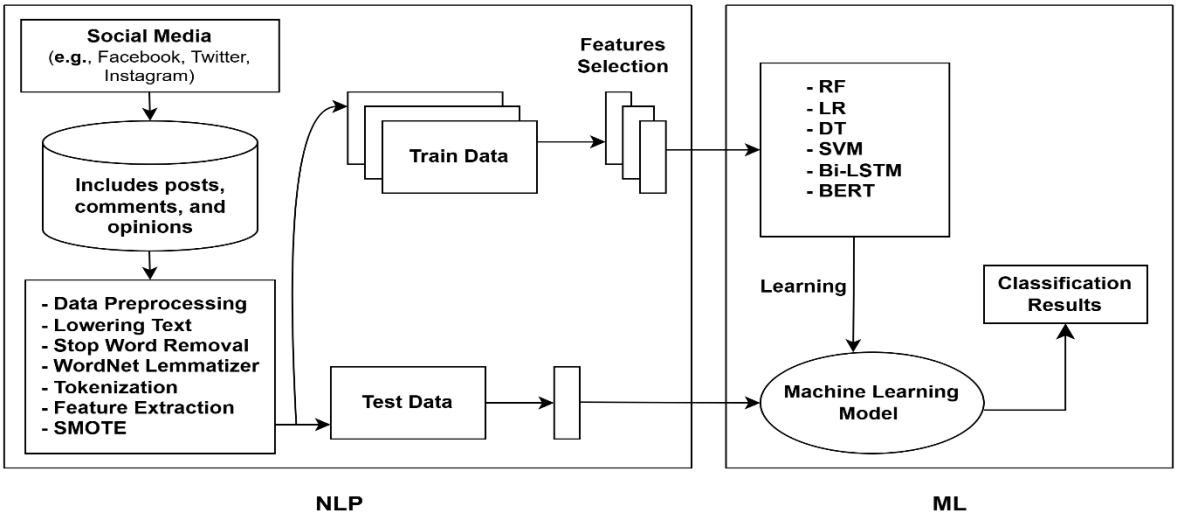

Fig. 1. The architecture diagram for social media cyberbullying detection

Creating a Model

There are two main phases to this approach: text analysis, and model selection.

1. Text Analysis

The proportion of pie chart posts that fit each cyberbullying categorization after cleaning the data is shown in Fig. 2 (Dhanalakshmi, et.al., 2023). Whereas religion has a very high number of entries 21.7% and not cyberbullying has a relatively low number of entries 16.7%, other categories have entries in about similar numbers.

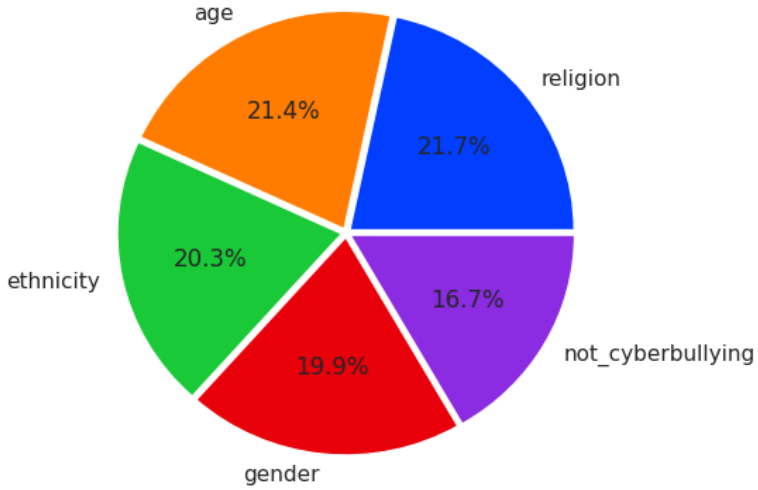

Fig. 2. Percentage of cyberbullying labels.

Fig. 3, illustrates the plot distribution where each cyberbullying label's sentence length is represented (Venu, et.al., 2023). Where gender contains short sentences while not cyberbullying contains long ones and others exhibit varying distributions of sentence lengths.

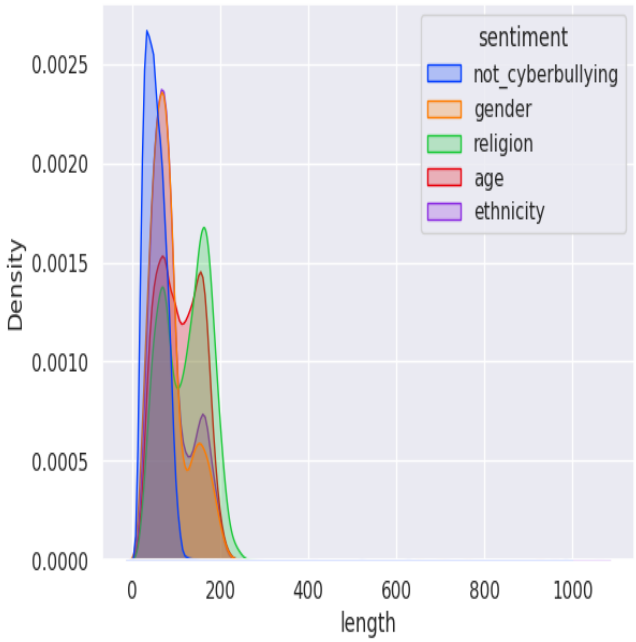

Fig. 3. Sentence length of each cyberbullying.

Fig. 4, illustrates the text length distribution by showing the number of texts for each length range, with the length measured in total word count. The majority of texts fall within the 2000 to 5000-word range, suggesting that this range may represent a standard or typical length for the dataset under study. Conversely, texts with fewer than 1000 words, and especially those below 200 in length, are notably rare. This skewed distribution highlights a potential preference for longer, more detailed texts in the analyzed context.

These findings are relevant because text length can influence both linguistic complexity and the depth of content analysis. The predominance of texts in the 2000–5000 words range aligns with the focus of this research on identifying patterns or trends in mid- to long-form texts. Moreover, the rarity of extremely short texts suggests that the dataset may not be well-suited for studies focusing on concise communications.

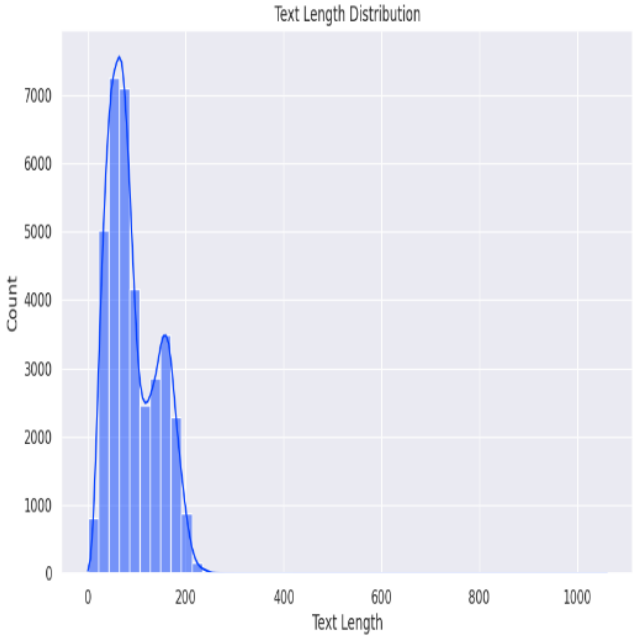

Fig. 4. Text length distribution of cyberbullying.

In Fig. 5, Polarity refers to the emotional tone of a text, which can range from negative to positive. Specifically, a polarity score between -1.0 and 0.0 indicates a negative emotional tone, while a score between 0.0 and 1.0 signifies a positive tone. Text with a score near 0.0 is considered neutral. Subjectivity, on the other hand, measures the degree to which a text expresses personal opinions or emotions as opposed to objective facts. A subjectivity score between 0.0 and 0.5 suggests a relatively factual and objective text, whereas scores between 0.5 and 1.0 indicate high subjectivity, reflecting a greater presence of opinions or emotional language.

In the context of cyberbullying, the data reveals that texts often have high subjectivity. This suggests that cyberbullying content frequently includes personal opinions, judgments, and emotional language, which are characteristics of subjective expression (**Van Hee, et.al., 2018**).

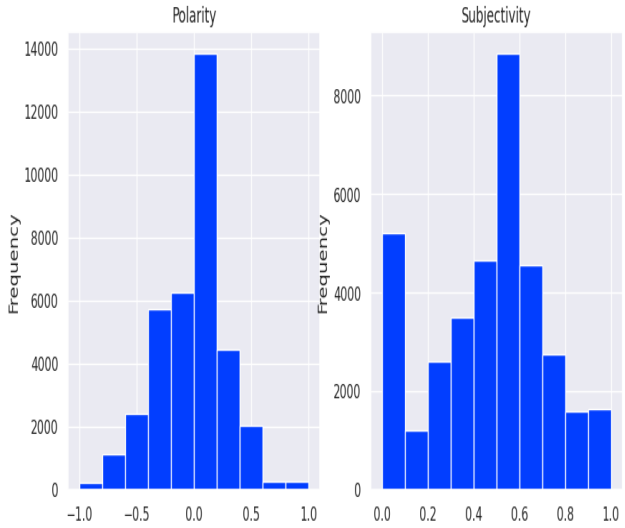

Fig. 5. Sentiment analysis plot of cyberbullying text.

Fig. 6, presents a word cloud highlighting the most commonly used terms associated with cyberbullying on social media. Larger words in the visualization signify higher frequency, indicating that these terms are prominent in online harassment. For instance, words like 'gay,' 'fuck,' 'rape,' and offensive slurs appear prominently, underscoring the use of derogatory or emotionally charged language. This word cloud is relevant to the findings as it provides insight into the linguistic patterns and vocabulary frequently observed in cyberbullying discourse, offering valuable cues for detection and mitigation strategies.

Fig. 6. Most common cyberbullying words

Model Selection

To automatically detect social media abuse, supervised multiclass classification ML algorithms such as RF, LR, DT, and SVM with radial basis kernel and DL algorithms such as Bi-LSTM and BERT are utilized (**Nemade, et.al., 2023**). The purpose of these models is to calculate the F1-score for each class (religion, age, ethnicity, gender, and not cyberbullying texts). These algorithms are used to classify the distinct categories of multiple classes. Tables 2 and 3 were employed to empirically evaluate and identify these machine-learning models. However, RF and BERT outperformed other models conducting nearly identical work. The classification report is evaluated as well. Additionally, the accuracy, precision, recall, f1-score, and are computed (**Dalvi, et.al., 2023**).

$$\text{Precision} = TP / (TP + FP)$$

$$\text{Recall} = TP / (TP + FN)$$

$$\text{F-Score} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$$

Table 2: Performance Analysis of Classification Algorithms

Criteria	RF	LR	DT	SVM
Type	Ensemble Learning	Statistical Method	Decision Tree	Margin-based Classifier
Use Case	Classification, Regression	Binary and Multiclass	Classification, Regression	Classification, Regression
Handling Overfitting	Mitigates Overfitting	Regularization Techniques	Prone to Overfitting	Controlled by Margin & C
Handling Outliers	Robust to Outliers	Sensitive to Outliers	Sensitive to Outliers	Sensitive to Outliers
Equations	Gini impurity, Entropy	Logistic Function	Gini impurity, Information Gain	Hyperplane, Kernel Trick
Multiclass Strategy	Majority Voting	One-vs-All	One-vs-Rest, One-vs-One	OvO, OvA, Direct-Multiclass
Implementation Ease	Easy with Libraries	Straight-forward	Straight-forward	Requires Parameter Tuning
Interpretability	Moderate	High	Moderate	Moderate to Low
Performance	Robust and Accurate	Good for Linear Problems	Prone to Overfitting	Effective in Complex-Spaces

Table 2 compares some of the most well-known ML algorithms that are frequently used for regression and classification problems, and Table III provides a comparison of several well-known DL algorithms. that are frequently used for classification problems. An informative summary of the key characteristics and performance attributes of DT, RF, LR, SVM, Bi-LSTM, and BERT is given in Tables 2, and 3. Overfitting management, outlier sensitivity, equations or concepts used, multiclass methods, ease of implementation, interpretability, and overall performance are some of the criteria used to evaluate each technique. This comparative

analysis assists researchers and practitioners in selecting the most suitable algorithm for our specific applications (Tahmasbi, et.al., 2018).

Table 3: Performance Analysis of Sequence Modeling Algorithms

Criteria	Bi-LSTM	BERT
Type	Recurrent Neural Network (RNN)	Transformer-based Pre-trained Model
Use Case	Sequential Data Processing (e.g., NLP tasks like Sentiment Analysis, Entity Recognition)	Contextual Language Understanding (e.g., Text Classification, Text Summarization)
Handling Long Dependencies	Good at capturing long-term dependencies in sequential data	Superior due to self-attention mechanisms for global context
Handling Outliers	Sensitive to Outliers	Robust due to large-scale pre-training
Criteria	**Bi-LSTM**	**BERT**
Training Paradigm	Requires task-specific training	Pre-trained on massive datasets; fine-tuning required for specific tasks
Equations	LSTM Cell Equations (Input, Forget, Output Gates)	Self-Attention, Positional Encoding
Multiclass Strategy	Task-Specific Output Layers (e.g., Softmax)	Pre-trained embeddings + Task-Specific Heads
Implementation Ease	Requires careful tuning of architecture and hyperparameters	Easy to fine-tune with pre-trained models
Interpretability	Moderate to Low	Low (interpreting attention weights requires additional techniques)
Performance	Effective for sequential tasks but slower compared to transformers	State-of-the-art across multiple NLP tasks; faster inference with optimized transformers

RESULT AND DISCUSSION

Here, we apply several ML techniques to an open-source Kaggle dataset of cyberbullying incidents in real-time, including RF, LR, DT, SVM, Bi-LSTM, and BERT (Kaggle, 2022). In order to prepare the dataset for training and testing, preprocessing, text analysis, and feature extraction are performed. The training dataset is oversampled using SMOTE for fairness, and it is split into a 20%-25% test size with a 40%-42% random state. The accuracy, f1-score, precision, and recall of these models are computed. Interestingly, RF and BERT did better than all the other methods combined (LR, DT, SVM, and Bi-LSTM).

The outcomes of setting up an experiment in which tweets are fetched from Twitter is now known as X using the Twitter API with the person's username and displaying the result of the fetched tweets, i.e., whether the tweets are bullying or not, along with their probability,

where from the 5th tweet is identified as "Bullying" and the remaining tweets are "Not-Bullying." In addition, the average probability of a bullied tweet is displayed at the end, down from 57% to 35%, meaning that the 5th tweet is classified as "Not-Bullying." The accuracy of RF, LR, DT, SVM, Bi-LSTM, and BERT are presented in Table IV. RF and BERT obtained the highest accuracy of 94%, followed by LR at 92%, DT at 92%, SVM at 93%, and Bi-LSTM achieved an accuracy of 93%.

Challenges in Real-Time Data Fetching Using APIs

While implementing real-time tweet analysis using the Twitter API (now known as X API), several challenges were encountered. These include:

1. Scalability
   Fetching tweets in real-time requires handling a potentially large influx of data, especially when monitoring popular or trending topics. This necessitates robust infrastructure to manage high throughput efficiently. In this study, we limited the scope to specific usernames to maintain manageable data loads. However, scaling this approach to include broader datasets or hashtags would require advanced strategies such as distributed computing or cloud-based solutions.
2. Computational Efficiency
   Preprocessing and classifying tweets in real-time impose a significant computational burden. Techniques like feature extraction and SMOTE-based oversampling add to the overhead. To address this, optimization methods such as batch processing and parallel execution could be explored in future work.
3. API Limitations
   The X API has inherent restrictions, such as rate limits that cap the number of requests allowed within a specific timeframe. Paid tiers of the API provide higher limits but increase operational costs. Additionally, the API's response time can vary based on network conditions, further complicating real-time analysis.

By considering these factors, we emphasize the need to design scalable and efficient systems that can handle real-time data processing while adhering to API limitations. Future iterations of this work may integrate strategies such as adaptive sampling, caching mechanisms, or transitioning to paid advanced tiers for extended access to enhance reliability and performance.

Tables 4, 5, 6, 7, 8, 9, and 10 indicate that the RF and BERT algorithms obtained the highest precision, while DT, SVM, LR, and Bi-LSTM achieved the lowest precision. In addition, RF, and BERT have attained superior recall and f1-score values compared to DT, SVM, LR, and Bi-LSTM

Table 4: The Accuracy of Classification Models

Classifiers	Accuracy
RF	94%
LR	92%
DT	92%
SVM	93%
Bi-LSTM	93%
BERT	94%

Table- 5: Classification Report of The RF

Category	Precision	Recall	F1-score
Religion	96%	96%	96%
Age	98%	97%	98%
Ethnicity	99%	99%	99%
Gender	96%	86%	91%
Not bullying	79%	89%	84%

Table 6: Classification Report of The LR

Category	Precision	Recall	F1-score
Religion	96%	94%	95%
Age	96%	96%	96%
Ethnicity	99%	98%	98%
Gender	96%	86%	90%
Not bullying	74%	87%	80%

Table 7: Classification Report of The DT

Category	Precision	Recall	F1-score
Religion	96%	92%	94%
Age	99%	96%	97%
Ethnicity	99%	98%	98%
Gender	95%	85%	88%
Not bullying	73%	91%	81%

Table- VIII: Classification Report of The SVM

Category	Precision	Recall	F1-score
Religion	97%	95%	96%
Age	95%	97%	96%
Ethnicity	98%	98%	98%
Gender	97%	86%	91%
Not bullying	75%	87%	81%

Table 9: Classification Report of The Bi-LSTM

Category	Precision	Recall	F1-score
Religion	95%	96%	96%
Age	98%	98%	98%
Ethnicity	98%	98%	98%
Gender	93%	85%	89%
Not bullying	79%	86%	82%

Table- X: Classification Report of The BERT

Category	Precision	Recall	F1-score
Religion	95%	98%	96%
Age	99%	98%	98%
Ethnicity	99%	99%	99%

Category	Precision	Recall	F1-score
Gender	92%	90%	91%
Not bullying	84%	84%	84%

CONCLUSION

Detecting and preventing social media cyberbullying using multiclass classification ML algorithms is a critical step toward addressing online abuse. Our proposed model was evaluated using RF, LR, DT, SVM, Bi-LSTM, and BERT algorithms, alongside feature extraction methods like TFIDF and CountVectorizer, and dataset balancing with SMOTE. The outcomes demonstrated that RF and BERT models achieved superior accuracy, with approximately 94% detection rates for identifying cyberbullying content. The study highlights the effectiveness of these models in safeguarding individuals from social media abuse.

Future efforts will focus on enhancing the proposed cyberbullying detection framework by integrating Vision-Language Models (VLMs) to analyze multimodal content like text, images, and memes. Scalability will be addressed through distributed computing and cloud-based solutions for real-time large-scale data processing, while computational efficiency will be improved using lightweight transformer architectures and adaptive sampling techniques. To overcome API limitations, strategies such as caching, request optimization, and transitioning to paid API advanced tiers will be employed. Additionally, multimodal fusion techniques will be expanded to incorporate audio and video analysis using models, creating a holistic detection system. A hybrid approach, combining traditional machine learning with advanced ensemble-based methods, will further enhance adaptability and robustness. Finally, emphasis will be placed on explainability and ethical considerations, ensuring transparent predictions while addressing bias and privacy concerns.

LIMITATIONS AND FUTURE RESEARCH RECOMMENDATIONS

While this study presents a robust framework integrating machine learning (ML) and deep learning (DL) models to detect and prevent cyberbullying across various social media platforms, several limitations must be acknowledged. *First,* the model evaluation relies primarily on a publicly available dataset from Kaggle, which, despite being well-labeled and balanced, may not capture the full complexity, linguistic diversity, or evolving nature of cyberbullying language in real-world online environments. The dataset's reliance on English-language content also restricts the applicability of the model in multilingual or non-Western digital contexts.

Second, although the framework incorporates multiple ML and DL algorithms—including Random Forest (RF), Logistic Regression (LR), Support Vector Machines (SVM), Bi-LSTM, and BERT—the performance assessments are limited to textual data only. Given the rise of multimodal cyberbullying involving images, memes, audio, and video content, the model's unidimensional focus may reduce its effectiveness in capturing the full spectrum of abusive online behavior. *Third,* the implementation of real-time analysis through social media APIs is constrained by practical limitations such as API rate limits, data access policies, and computational costs. These technical and financial barriers could hinder the framework's scalability, especially when deployed in large-scale monitoring systems or across multiple platforms simultaneously. *Fourth,* while the framework performs exceptionally well in terms of accuracy and F1-scores—particularly with RF and BERT classifiers—it lacks comprehensive

exploration of ethical considerations such as data privacy, false positives, model explainability, and the psychological impact of automated moderation on users. These dimensions are crucial in the design of responsible AI systems for sensitive contexts like cyberbullying detection.

In light of these limitations, several directions are proposed for future research:

1. Expansion to Multilingual and Multicultural Contexts
   Future studies should develop and evaluate datasets encompassing various languages, dialects, and regional discourses to enhance the generalizability of detection models across global platforms.
2. Integration of Multimodal Data
   Subsequent research should incorporate image, video, and audio processing capabilities using vision-language models (VLMs) and multimodal fusion techniques to address cyberbullying in its diverse forms.
3. Optimization for Real-Time Performance
   Further work is needed to optimize model efficiency through lightweight architectures, ensemble learning, and distributed computing, ensuring scalability and responsiveness without compromising accuracy.
4. Addressing Ethical and Legal Dimensions
   Future efforts must embed ethical principles such as fairness, accountability, and transparency into model development. This includes the implementation of explainable AI (XAI) methods, user consent mechanisms, and robust data anonymization techniques to align with legal standards like GDPR.
5. User-Centric Evaluation and Feedback Loops
   Incorporating user feedback into the classification process and assessing the social impact of automated moderation tools will enhance model adaptability and legitimacy, fostering safer digital communities through participatory design.

REFERENCES

Al-Dabet, S., ElMassry, A., Alomar, B., & Alshamsi, A. (2023, March). Transformer-based Arabic offensive speech detection. 2021 *International Conference on Emerging Smart Computing and Informatics (ESCI)*, 1–6. https://doi.org/10.1109/esci56872.2023.10100134

Alothali, E., Alashwal, H., Salih, M., & Hayawi, K. (2021, October). Real-time detection of social bots on Twitter using machine learning and Apache Kafka. *2021 5th Cyber Security in Networking Conference (CSNet)*. https://doi.org/10.1109/csnet52717.2021.9614282

Balet, T., Vo, Q., Salem, O., & Mehaoua, A. (2023, June). Cyberbullying detection on tweets from Twitter using machine learning algorithms. *2023 International Conference on Intelligent Computing, Communication, Networking and Services (ICCNS)*, 3, 177–182. https://doi.org/10.1109/iccns58795.2023.10193450

Beniwal, R., Jha, S., Mehta, S., & Dhiman, R. (2023, June). 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

Cyberbullying classification. (2022, January 17). *Kaggle*. https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification

Dalvi, R. R., Chavan, S. B., & Halbe, A. (2020, May). Detecting a Twitter cyberbullying using machine learning. *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)*. https://doi.org/10.1109/iciccs48265.2020.9120893

Dhanalakshmi, D., Rani, N. D., Pendam, K., Hariharan, S., Kukreja, V., & Jayakshata, P. (2023, June). Machine learning based intelligent cyberbullying avoidance system. *2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, 9, 1594–1597. https://doi.org/10.1109/icscss57650.2023.10169376

Gupta, S., Vadgama, U., & Vedhavathy, T. R. (2023, April). Identification and labeling of textual cyberbullying using BiLSTM and BERT. *2023 International Conference on Networking and Communications (ICNWC)*, 1–5. https://doi.org/10.1109/icnwc57852.2023.10127308

Islam, M., Rahman, M., Ahmed, M. T., Islam, A. Z. M., Das, D., & Hoque, M. M. (2023, February). Sexual harassment detection using machine learning and deep learning techniques for Bangla text. *2023 International Conference on Electrical, Computer and Communication Engineering (ECCE)*. https://doi.org/10.1109/ecce57851.2023.10101522

Jeevitha, R., Chaitanya, K., Mathesh, N., Nithyanarayanan, B., & Darshan, P. (2023, March). Using machine learning to identify instances of cyberbullying on social media. *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, 8, 207–212. https://doi.org/10.1109/icscds56580.2023.10104720

Matias, S. M. M., Costales, J. A., & De Los Santos, C. M. (2022, August). A framework for cybercrime prediction on Twitter tweets using text-based machine learning algorithm. *2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)*, 10, 235–240. https://doi.org/10.1109/prai55851.2022.9904212

Nemade, S., Mane, S. B., & Nandgaonkar, S. (2023, April). Detection and classification of aggressive comments and hate speech. *2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES)*, 14, 55–60. https://doi.org/10.1109/cises58720.2023.10183445

Patidar, M. (2021, November). Cyber bullying detection for Twitter using ML classification algorithms. *International Journal for Research in Applied Science and Engineering Technology*, 9(11), 24–29. https://doi.org/10.22214/ijraset.2021.38701

Pawar, V., Jose, D. V., & Patil, A. (2022, December). Explainable AI method for cyber bullying detection. *2022 IEEE 2nd International Conference on Mobile Networks and Wireless Communications (ICMNWC)*. https://doi.org/10.1109/icmnwc56175.2022.10031652

Perez, C., & Karmakar, S. (2023, March). An NLP-assisted Bayesian time-series analysis for prevalence of Twitter cyberbullying during the COVID-19 pandemic. *Social Network Analysis and Mining*, 13(1). https://doi.org/10.1007/s13278-023-01053-4

Pericherla, S., & Ilavarasan, E. (2020). A study of machine learning approaches to detect cyberbullying. *In Lecture Notes in Networks and Systems* (pp. 369–377). https://doi.org/10.1007/978-981-15-5397-4_38

Pradhan, A., Yatam, V. M., & Bera, P. (2020, June). Self-attention for cyberbullying detection. *2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)*. https://doi.org/10.1109/cybersa49311.2020.9139711

Rachidi, R., Ouassil, M. A., Errami, M., Cherradi, B., Hamida, S., & Silkan, H. (2023, May). Social media's toxic comments detection using artificial intelligence techniques. *2023 3rd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*. https://doi.org/10.1109/iraset57153.2023.10153015

Raj, M., Singh, S., Solanki, K., & Selvanambi, R. (2022, July). An application to detect cyberbullying using machine learning and deep learning techniques. *SN Computer Science*, 3(5). https://doi.org/10.1007/s42979-022-01308-5

Romsaiyud, W., Nakornphanom, K. N., Prasertsilp, P., Nurarak, P., & Konglerd, P. (2017, February). Automated cyberbullying detection using clustering appearance patterns. *2017 9th International Conference on Knowledge and Smart Technology (KST)*, 242–247. https://doi.org/10.1109/kst.2017.7886127

Sainju, K. D., Mishra, N., Kuffour, A., & Young, L. (2021, February). Bullying discourse on Twitter: An examination of bully-related tweets using supervised machine learning. *Computers in Human Behavior*, 120, 106735. https://doi.org/10.1016/j.chb.2021.106735

Selvaraj, P. A., Jagadeesan, M., Saravanan, T. M., Shaheela, Y., Pandey, S. K., & Sowmithra, M. (2023, January). Detection of personal attacks in mobbing on social media using deep learning techniques. *2022 International Conference on Computer Communication and Informatics (ICCCI)*. https://doi.org/10.1109/iccci56745.2023.10128345

Shanto, S. B., Islam, M. J., & Samad, M. A. (2023, February). Cyberbullying detection using deep learning techniques on Bangla Facebook comments. *2023 International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC)*, 1–7. https://doi.org/10.1109/isacc56298.2023.10083690

Tahmasbi, N., & Rastegari, E. (2018, December). A socio-contextual approach in automated detection of public cyberbullying on Twitter. *ACM Transactions on Social Computing*, 1(4), 1–22. https://doi.org/10.1145/3290838

Thangarasu, G., & Alla, K. R. (2023, May). Detection of cyberbullying tweets in Twitter media using random forest classification. *2023 IEEE 13th Symposium on Computer Applications & Industrial Electronics (ISCAIE)*. https://doi.org/10.1109/iscaie57739.2023.10165118

Van Hee, C., et al. (2018, October). Automatic detection of cyberbullying in social media text. *PLOS ONE*, 13(10), e0203794. https://doi.org/10.1371/journal.pone.0203794

Venu, V. S., Shanmugasundaram, H., Seelam, M. R., Kotha, V. V. R., Muthyala, S. S. R., & Kansal, S. (2023, May). Detection of cyberbullying on user tweets and Wikipedia text using machine learning. *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, 12, 327–332. https://doi.org/10.1109/icaaic56838.2023.10140252

Wang, J., Fu, K., & Lu, C.-T. (2020, December). SOSNET: A graph convolutional network approach to fine-grained cyberbullying detection. *2021 IEEE International Conference on Big Data (Big Data)*. https://doi.org/10.1109/bigdata50022.2020.9378065

Zhang, J., Otomo, T., Li, L., & Nakajima, S. (2019, October). Cyberbullying detection on Twitter using multiple textual features. *2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST)*, 1–6. https://doi.org/10.1109/icawst.2019.8923186