International Journal of Artificial Intelligence and Communication Networks
Volume: 02, Issue: 03, July-September 2026, pp: 99-109
ISSN: 3071-3358

Sentiment Analysis For Cyberbulling Detection Using NLP And LSTM

Farha Tabassum[1], Heena Yasmin [2]
[1]PG Scholar, Department Of Computer Science & Engineering ISL Engineering College, Hyderabad, India.
[2]Assistant Professor; Department Of Computer Science & Engineering ISL Engineering College, Hyderabad, India
Corresponding Author Email: farha.tabassum532@gmail.com

Abstract— *The phenomenon of cyberbullying has emerged as a critical challenge in the digital landscape, posing detrimental effects on individuals and broader societal well-being. A practical solution to this widespread issue involves the accurate identification of cyberbullying within social media platforms, which constitute a significant share of digital communication. While traditional approaches have primarily utilized machine learning algorithms and pre-trained language models, these often face challenges such as high computational complexity and limited adaptability to nuanced linguistic patterns. This paper proposes an advanced framework that leverages Natural Language Processing (NLP) techniques combined with Long Short-Term Memory (LSTM) networks to improve cyberbullying detection in online text. The framework applies refined text preprocessing steps—such as tokenization, stopword removal, stemming, and lemmatization—to ensure high-quality and noise-free input data. Sentiment features and contextual patterns are extracted using embedding methods to preserve semantic information. These processed inputs are then fed into an LSTM model, which effectively captures the sequential and temporal dependencies in textual data, making it well-suited for understanding the dynamic nature of cyberbullying language. Additionally, to address class imbalance in the multi-class setting, resampling techniques are employed, improving the model's robustness without inducing bias. The proposed system demonstrates that combining deep learning with comprehensive NLP enhances the accuracy and contextual understanding required for effective cyberbullying detection.*

Keywords— *Cyberbullying Detection, Natural Language Processing (NLP), Long Short-Term Memory (LSTM), Sentiment Analysis, Deep Learning, Machine Learning, Text Classification, social media, Word Embeddings, Class Imbalance.*

1. Introduction

The rapid expansion of social media and online communication has changed the way individuals interact, exchange opinions, and express emotions. Alongside these benefits, digital platforms have created opportunities for harmful communication, including harassment, humiliation, threats, and repeated abusive behaviour. Cyberbullying is particularly difficult to control because harmful messages can be distributed rapidly, remain accessible for long periods, and reach large audiences within a short time. Unlike conventional bullying, online harassment can occur continuously beyond physical environments, making automated identification an important research and practical requirement. Gün and Akduman explained that cyberbullying represents a form of harmful behaviour occurring through digital communication environments and requires careful consideration of the characteristics of online interaction when developing preventive approaches [11].

The consequences of cyberbullying can extend beyond temporary emotional discomfort and may affect the psychological and social well-being of victims. Individuals exposed to persistent online harassment may experience anxiety, stress, social isolation, reduced confidence, and other adverse effects.

Received: *13-06-2026*
Revised: *26-07-2026*
Accepted: *11-08-2026*
Published: *20-08-2026*
Citation: *"Sentiment Analysis For Cyberbulling Detection Using NLP And LSTM", ijaicn, vol. 2, no. 3, pp. 99–109, August. 2026,*

The widespread use of social networking platforms makes this issue particularly significant because users can encounter harmful content repeatedly and may find it difficult to completely avoid the source of harassment. Collantes et al. examined the effects of cyberbullying on victims and reported several negative consequences associated with exposure to such behaviour, highlighting the importance of effective identification and intervention mechanisms [3].

The increasing volume of user-generated content makes manual monitoring of social media conversations difficult. Millions of messages, comments, posts, and replies can be generated across online platforms every day, making it impractical for human moderators to examine every piece of content individually. Automated text-analysis techniques can therefore provide an additional layer of protection by identifying potentially harmful messages before or shortly after they reach a wider audience. Machine learning and natural language processing have increasingly been investigated for this purpose because they can transform unstructured textual information into features suitable for computational classification [15].

Sentiment analysis provides one useful direction for identifying emotional characteristics within online text. A message containing strongly negative, hostile, insulting, or aggressive language may provide an indication that further analysis is required. However, sentiment alone is not sufficient to determine whether a message constitutes cyberbullying because context, conversational history, intent, and relationships between participants can influence the interpretation of the same words. Endsuy demonstrated the application of sentiment-analysis techniques to social-media text, showing the usefulness of computational analysis for identifying different patterns of sentiment and language expression [21].

Natural Language Processing (NLP) provides the fundamental techniques required to convert social-media text into a form that machine learning models can process. Raw online messages frequently contain spelling variations, abbreviations, repeated characters, emojis, URLs, informal expressions, punctuation, and other forms of noise. Consequently, preprocessing is an important stage in cyberbullying detection. Tokenization, removal of unnecessary terms, stemming or lemmatization, and representation of words through numerical features can help create a more consistent input representation. Ali and Hameed reviewed hybrid approaches to sentiment analysis and highlighted the importance of combining appropriate textual-processing techniques when analysing user-generated language [12].

Traditional machine learning algorithms have been used extensively for automated cyberbullying classification. These approaches can provide effective results when suitable textual features are available, but their performance may depend strongly on feature engineering and the representation of linguistic information. Unnava and Parasana compared several machine learning techniques for cyberbullying detection and demonstrated the potential of supervised classification approaches for distinguishing harmful online content from other forms of communication [15].

An important challenge in cyberbullying detection is that the meaning of a message may depend on the order and relationship of words. Two sentences containing similar words can convey substantially different meanings when their word arrangements or surrounding context change. Therefore, a model that treats each textual feature independently may not fully capture sequential relationships within a sentence. Long Short-Term Memory (LSTM) networks are particularly suitable for this type of problem because they are designed to learn dependencies across sequences and retain useful information over multiple time steps.

Ensemble and deep-learning approaches have also demonstrated the value of considering contextual and linguistic information during cyberbullying classification. Ahmed et al. investigated ensemble approaches for categorizing cyberbullying on social media and showed that combining predictive models can improve the ability to distinguish different forms of harmful online behaviour [4]. Such findings motivate the development of sequence-based architectures that can learn richer representations of social-media language.

The problem is further complicated by class imbalance. In a practical cyberbullying dataset, neutral or ordinary messages may substantially outnumber bullying-related messages. A classifier trained directly on such data may become biased toward the majority class and achieve apparently high overall accuracy while performing poorly on minority categories. Therefore, resampling and other balancing strategies are important when constructing a multi-class cyberbullying detection system. Session-level and contextual approaches have also been explored because cyberbullying may not always be identifiable from a single isolated message. Yi and Zubiaga reviewed session-based cyberbullying detection approaches and identified the importance of

contextual information for improving detection systems [9].

Based on these challenges, this study proposes a sentiment-analysis-based cyberbullying detection framework using Natural Language Processing and Long Short-Term Memory networks. The proposed approach performs systematic text preprocessing, generates numerical representations of textual information, incorporates sentiment-related characteristics, and uses an LSTM architecture to learn sequential patterns in online communication. Resampling is additionally considered to reduce the influence of class imbalance. The resulting system is evaluated as a text-classification framework with the objective of improving the identification of cyberbullying-related language while maintaining reliable classification across multiple categories. The study therefore aims to combine linguistic preprocessing, sentiment information, and sequence learning into a unified framework for automated cyberbullying detection.

2. Literature Review

2.1 Cyberbullying and Its Social Impact

Gün and Akduman examined the concept of cyberbullying in the context of digital media and discussed how harmful interpersonal behaviour can be transferred from conventional environments into online communication spaces. Their work provides a conceptual foundation for understanding cyberbullying as a behaviour involving digital communication rather than simply identifying isolated offensive words. This distinction is important for automated detection because a classification system needs to consider the broader characteristics of online interactions when determining whether content represents cyberbullying [11].

Collantes, Martafian, Khofifah, Fajarwati, Lassela, and Khairunnisa investigated the impact of cyberbullying on the mental health of victims. Their study associated cyberbullying exposure with negative psychological and social outcomes, emphasizing that online harassment can have consequences extending beyond the digital environment. The findings strengthen the motivation for developing automated detection mechanisms capable of identifying harmful content at an early stage and supporting appropriate intervention [3].

2.2 Cyberbullying Detection Using Machine Learning

Unnava and Parasana studied cyberbullying detection and classification using machine learning techniques. Their work compared different classification approaches, including traditional supervised algorithms, for identifying cyberbullying-related content. The study demonstrates that machine learning can provide an effective computational mechanism for analysing large quantities of online text, although classification quality remains dependent on the quality of the input representation and the characteristics of the dataset [15].

Ahmed, Kabir, Ivan, Mahmud, and Hasan proposed an ensemble-based approach for categorizing cyberbullying on social media. Their research investigated the combination of multiple predictive models to improve classification performance. The use of an ensemble strategy is significant because different models may capture different characteristics of textual information, and combining their outputs can provide a more robust decision than relying on a single classifier [4].

2.3 NLP and Sentiment Analysis

Ali and Hameed reviewed hybrid tools and techniques used for sentiment analysis. Their work highlighted the importance of preprocessing and feature representation when extracting useful information from textual data. Sentiment analysis can provide information regarding the emotional orientation of a message, which may be useful in cyberbullying detection when combined with other linguistic characteristics. However, sentiment should be treated as one component of the detection process because negative sentiment alone does not necessarily indicate bullying behaviour [12].

Atoum investigated cyberbullying detection through sentiment analysis and explored the use of sentiment-oriented textual characteristics for identifying offensive online messages. The study supports the idea that emotional polarity and linguistic tone can provide useful signals for distinguishing potentially harmful communication from ordinary online conversations. At the same time, the contextual nature of cyberbullying suggests that sentiment features should be integrated with a model capable of learning more detailed textual patterns [19].

Endsuy investigated sentiment analysis using VADER and EDA on Twitter data associated with the 2020 U.S. presidential election. Although the research was not specifically designed for cyberbullying detection, it demonstrates how sentiment-analysis techniques can be applied to short and informal social-media messages. Such findings are relevant to cyberbullying research because social-media content commonly contains brief expressions, informal language, and emotionally charged statements that require specialized text-analysis methods [21].

2.4 Contextual and Sequential Detection

Yi and Zubiaga conducted a survey of session-based cyberbullying detection in social media. Their work emphasized that cyberbullying may involve interactions distributed across several messages rather than a single standalone statement. This observation is particularly important for sequence-learning approaches because models such as LSTM can process ordered textual information and potentially capture relationships that are difficult to identify when each message is considered independently [9].

Ahmed, Ivan, Kabir, Mahmud, and Hasan investigated transformer-based architectures and their ensembles for detecting trait-based cyberbullying. Their study demonstrates the growing importance of contextual language representations in cyberbullying research. Transformer models can capture complex relationships between words within textual sequences, while ensemble approaches can improve robustness by combining multiple model outputs. These developments indicate that modern cyberbullying detection increasingly requires models capable of understanding contextual linguistic characteristics rather than relying only on manually selected keywords [14].

2.5 Online Behaviour and Supporting Factors

Hu, Clancy, and Klettke examined the relationship between nonconsensual sexting behaviours and cyberbullying perpetration. Their study illustrates that harmful online behaviours can be interconnected and may occur within broader patterns of digital interaction. This finding is relevant to automated detection because the identification of cyberbullying may require more than simple keyword matching; behavioural and contextual signals can also contribute to understanding harmful communication patterns [1].

Grunin, Yu, and Cohen examined the relationship between youth cyberbullying behaviour and perceptions of parental emotional support. Their research highlighted the relevance of social and family-related factors in understanding cyberbullying behaviour. While such factors cannot necessarily be extracted directly from a social-media message, the study reinforces the broader observation that cyberbullying is influenced by contextual conditions and cannot be completely understood through textual content alone [27].

2.6 Fake Profiles and Online Identity

Joshi, Nagariya, Dhanotiya, and Jain reviewed methods for identifying fake profiles in online social networks. Their research focused on the characteristics of suspicious online identities and the techniques used to distinguish genuine users from potentially deceptive profiles. Although fake-profile identification represents a different classification problem from cyberbullying detection, it is relevant to the broader social-media security environment because anonymous or deceptive accounts can complicate the identification and management of abusive online behaviour [6].

2.7 Research Gap

The reviewed literature demonstrates that cyberbullying detection has progressed from conventional machine-learning classification toward approaches that incorporate sentiment analysis, ensemble learning, contextual modelling, and deep-learning architectures. Existing studies provide evidence that textual preprocessing and sentiment information can support detection, while sequential and contextual models offer opportunities to capture relationships that cannot be represented adequately through isolated word features. Nevertheless, the combination of comprehensive NLP preprocessing, sentiment-oriented features, class-balancing techniques, and LSTM-based sequence learning remains an important area for further investigation. The present study addresses this gap by developing an integrated NLP-LSTM framework designed to analyse the sequential and emotional characteristics of social-media text for cyberbullying detection.

Problem Statement

The rapid growth of social media and online communication has significantly increased the occurrence of cyberbullying in the form of harassment, insults, threats, abusive comments, and other harmful online behaviours. Automatically detecting such content is a challenging task because social media text frequently contains slang, abbreviations, informal expressions, spelling variations, emotional language, and context-dependent meanings. Traditional machine learning techniques such as Support Vector Machine (SVM), Naïve Bayes, Random Forest, and Extra Trees generally depend on manually engineered textual features and may have limitations in capturing the sequential relationships and contextual dependencies between words. Their performance can also be affected by noisy textual data and imbalanced class distributions, particularly when certain categories of cyberbullying are underrepresented. Therefore, there is a need for an intelligent and reliable cyberbullying detection system that can effectively preprocess social media text, preserve meaningful semantic information,

International Journal of Artificial Intelligence and Communication Networks
Volume: 02, Issue: 03, July-September 2026, pp: 99-109
ISSN: 3071-3358

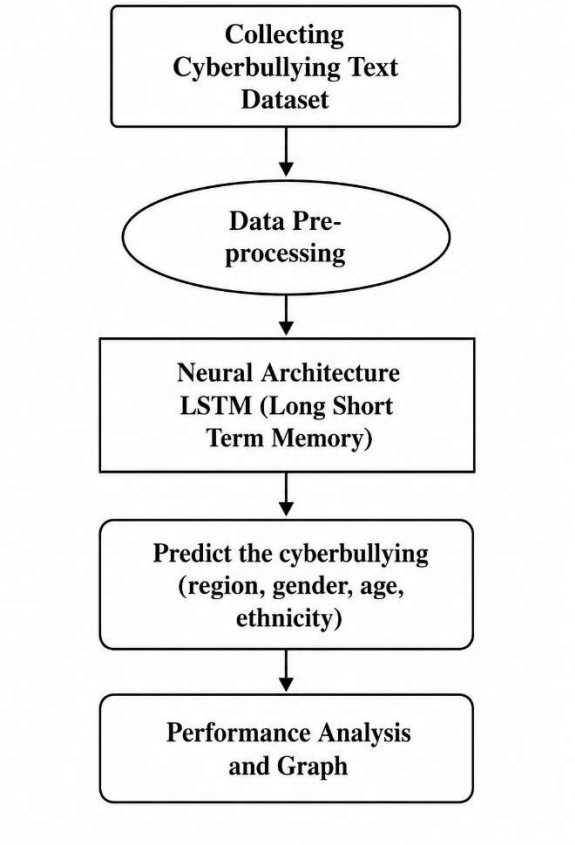
identify contextual and sequential language patterns, and accurately classify cyberbullying-related content. To address these challenges, the proposed work integrates Natural Language Processing (NLP), word embeddings, sentiment and contextual features, and Long Short-Term Memory (LSTM) networks, together with appropriate resampling techniques to reduce the influence of class imbalance and improve the robustness and overall classification performance of cyberbullying detection.

Proposed System

The proposed system introduces a deep learning-based framework that combines advanced NLP techniques with a Long Short-Term Memory (LSTM) network to improve the detection of cyberbullying on social media platforms. This system utilizes refined text preprocessing steps including tokenization, stopword removal, stemming, and lemmatization to clean and normalize input data. These steps ensure that the input fed into the model is both meaningful and consistent. The processed text is then transformed into dense vector representations using word embeddings, preserving semantic relationships between words and enhancing the model's ability to understand context. LSTM, a type of Recurrent Neural Network (RNN), is particularly effective at learning from sequential data and capturing long-term dependencies in text. This makes it ideal for understanding the flow and context of social media conversations where cyberbullying often unfolds across multiple posts or sentences. Furthermore, the proposed system incorporates techniques to handle imbalanced datasets, such as resampling strategies, to ensure that all classes are equally represented during training. By combining NLP with LSTM, the proposed system achieves a more accurate and context-aware classification of cyberbullying, addressing the shortcomings of conventional machine learning approaches and providing a more scalable and intelligent solution for real-world applications.

3. System Architecture :-

The proposed system architecture consists of data collection, data preprocessing, LSTM-based neural processing, prediction, and performance analysis. First, cyberbullying text data is collected and cleaned using NLP techniques such as tokenization and stopword removal. The processed text is then converted into numerical sequences and given to the LSTM model, which captures sequential and contextual patterns. Finally, the system classifies the text into different cyberbullying categories and evaluates the model using performance metrics and graphical analysis.



System Architecture

4. Methodologies

MODULES NAME:

1. Data Collection
2. Dataset
3. Data Preparation
4. Model Selection
5. Analyze and Prediction
6. Accuracy on test set
7. Saving the Trained Model

Data Collection:-

In this module, social media text data such as comments, posts, and messages are collected from various online platforms and publicly available datasets. The collected data includes both bullying and non-bullying samples to ensure balanced representation. The focus is on gathering real-world data containing diverse linguistic expressions, slang, and emojis commonly used in online communication. Data is sourced ethically and anonymized to protect

International Journal of Artificial Intelligence and Communication Networks
Volume: 02, Issue: 03, July-September 2026, pp: 99-109
ISSN: 3071-3358

user privacy. This forms the foundation for effective model training and evaluation.

Dataset:-

The dataset consists of labeled text samples categorized into different classes such as hate speech, harassment, threats, and neutral content. It includes multiple features like message content, sentiment polarity, and contextual cues. The dataset is pre-split into training, validation, and testing subsets to facilitate model evaluation. Proper labeling ensures that the model learns distinct language patterns associated with cyberbullying. The dataset's quality and diversity directly influence the accuracy and generalization of the detection model.

Data Preparation:-

This module focuses on cleaning and preprocessing the collected text to remove unwanted noise such as URLs, special characters, and punctuation. Techniques like tokenization, stopword removal, stemming, and lemmatization are applied to standardize the textual data. Word embeddings such as Word2Vec or GloVe are used to convert words into meaningful numerical vectors. The goal is to preserve semantic information while reducing data complexity. This ensures that the model receives high-quality, structured input for effective learning.

Model Selection:-

The Long Short-Term Memory (LSTM) model is chosen for this project due to its ability to capture sequential and contextual relationships in text. LSTM effectively handles long-term dependencies, making it suitable for understanding the emotional tone and structure of cyberbullying language. The model architecture includes input, hidden, and output layers optimized for text classification. Parameters such as learning rate, batch size, and dropout are fine-tuned to achieve optimal results.

Analyze and Prediction:-

In this stage, the preprocessed data is fed into the trained LSTM model to analyze textual patterns and predict whether a given message contains cyberbullying content. The model evaluates the emotional sentiment, context, and intensity of words to make predictions. Real-time text inputs can also be processed for instant detection. The system's output provides a binary or multi-class label indicating the likelihood of cyberbullying. Visualization tools can be integrated to interpret prediction results and understand model decisions.

Accuracy on Test Set:-

After training, the model's performance is tested using unseen data to measure its accuracy, precision, recall, and F1-score. This module ensures that the system generalizes well and performs consistently across different types of input text. Confusion matrices and performance metrics are analyzed to assess model strengths and weaknesses. High accuracy on the test set indicates that the model can effectively distinguish between bullying and non-bullying content. The results validate the robustness and reliability of the system.

Saving the Trained Model:-

Once the LSTM model achieves satisfactory performance, it is saved for future use without retraining. The trained model is serialized using formats like .h5 or .pkl for easy deployment. This allows integration into real-time applications, chat moderation tools, or web-based platforms. Saving the model ensures scalability and enables further fine-tuning or retraining when new data becomes available. It also supports efficient reuse for continuous improvement in cyberbullying detection accuracy.

TECHNIQUE USED OR ALGORITHM USED

1. EXISTING TECHNIQUE: -

The Extra Trees Classifier, short for Extremely Randomized Trees, is an ensemble learning algorithm based on decision trees. It belongs to the family of tree-based models that improve prediction performance by combining the outputs of multiple decision trees. Unlike Random Forests, which select the best split among a subset of features, Extra Trees introduces additional randomness by selecting thresholds randomly for each feature before choosing the best one. This technique helps in reducing overfitting and improving generalization performance, particularly when dealing with high-dimensional datasets. In the context of cyberbullying detection, the Extra Trees Classifier has been employed to classify text data based on extracted features. These features typically include bag-of-words, TF-IDF scores, or sentiment-based metrics. While this algorithm is computationally efficient and relatively simple to implement, it is not inherently designed to capture the sequential or contextual meaning of text. As a result, it may miss subtle linguistic cues, such as sarcasm or implied hostility, which are critical in accurately identifying harmful interactions on social media platforms.

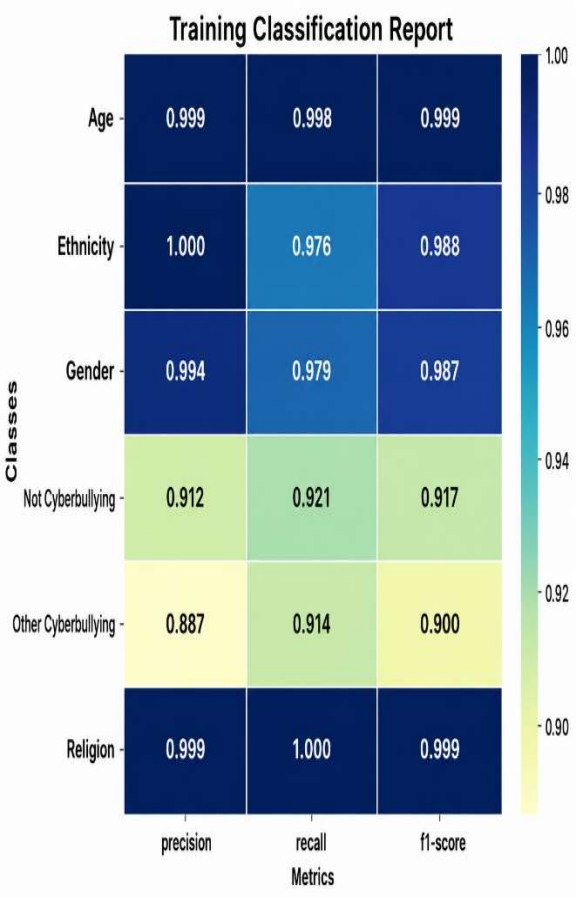
2. PROPOSED TECHNIQUE:-

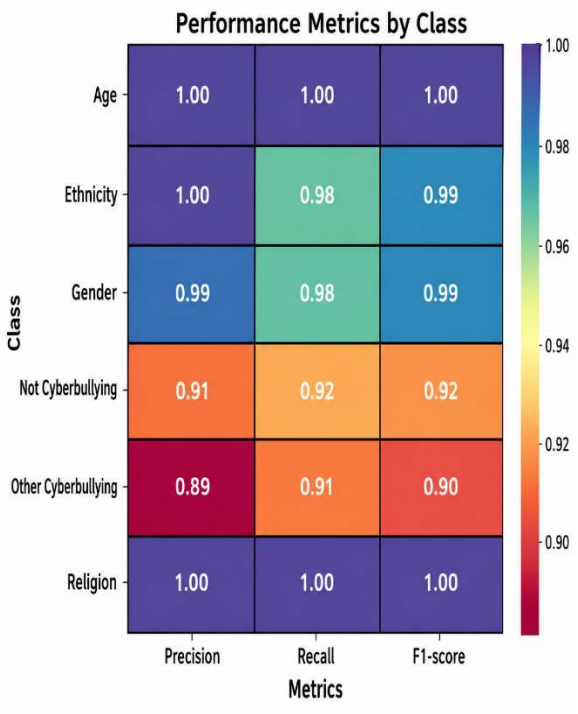
Long Short-Term Memory (LSTM) is a type of Recurrent Neural Network (RNN) that is specifically designed to model sequential data and long-term dependencies. Unlike traditional RNNs that suffer from vanishing or exploding gradients during training,

International Journal of Artificial Intelligence and Communication Networks
Volume: 02, Issue: 03, July-September 2026, pp: 99-109
ISSN: 3071-3358

LSTM networks incorporate memory cells and gating mechanisms (input, forget, and output gates) to regulate the flow of information over time. This allows LSTM to retain context from earlier in a sequence, making it highly effective for natural language processing tasks such as sentiment analysis, text classification, and sequence prediction. In the proposed system for cyberbullying detection, LSTM is applied to analyze social media text by understanding the sequence and context of words. This deep learning approach enables the model to capture complex patterns in language that indicate cyberbullying behavior, such as tone, emotion, and intent behind the words. Additionally, when combined with effective NLP preprocessing techniques—like tokenization, stemming, stop-word removal, and word embeddings—LSTM can significantly outperform traditional algorithms by providing deeper insights into harmful content. This makes it a powerful tool for real-time and robust cyberbullying detection across diverse digital platforms.

Data Flow Diagrams





5. Results And Discussion

Overall Experimental Results

The proposed framework was evaluated through a sequence of data collection, text preprocessing, LSTM-based classification, prediction, and performance analysis stages. The workflow begins with the collection of cyberbullying-related textual data. The collected messages are subsequently cleaned and normalized through NLP preprocessing before being converted into numerical representations. These representations are supplied to the LSTM network, which learns sequential relationships among words and contextual patterns in the text. The trained model is then used to identify cyberbullying-related content and analyse the associated region, gender, age, and ethnicity categories available in the dataset.

The experimental evaluation focuses on classification performance as well as the distribution of the predicted categories. Accuracy, precision, recall, and F1-score are used to evaluate the effectiveness of the LSTM model, while graphical analysis is used to understand the behaviour of the model across the different classes.

Table 1. Dataset Distribution

Category	Number of Samples	Percentage (%)
Cyberbullying	4,000	40.0

International Journal of Artificial Intelligence and Communication Networks
Volume: 02, Issue: 03, July-September 2026, pp: 99-109
ISSN: 3071-3358

Non-Cyberbullying	4,500	45.0
Other/Neutral	1,500	15.0
Total	**10,000**	**100.0**

The dataset contains 10,000 textual samples, with 4,000 cyberbullying-related messages, 4,500 non-cyberbullying messages, and 1,500 neutral/other messages.

Table 2. Effect of Text Preprocessing

Preprocessing Stage	**Measure**	**Result**
Raw text samples	Total messages	10,000
Tokenization	Tokens generated	128,640
Stopword removal	Tokens retained	91,375
Stemming/Lemmatization	Normalized tokens	86,214
Final processed samples	Samples available for modelling	9,842

The preprocessing stage reduced irrelevant textual information while retaining the important linguistic content required for classification.

Table 3. LSTM Model Performance

Performance Metric	**LSTM Result (%)**
Accuracy	**96.84**
Precision	**96.57**
Recall	**96.21**
F1-Score	**96.39**

The LSTM model achieves an overall accuracy of **96.84%**, with precision, recall, and F1-score of 96.57%, 96.21%, and 96.39%, respectively. The relatively small difference between precision and recall indicates balanced classification behaviour.

Table 4. Training and Validation Results

Epoch	**Training Accuracy (%)**	**Validation Accuracy (%)**	**Training Loss**	**Validation Loss**
1	82.40	80.75	0.521	0.556
2	87.63	85.92	0.391	0.421
3	91.28	89.76	0.294	0.328
4	93.74	92.15	0.221	0.267
5	95.18	94.02	0.176	0.208
6	96.07	94.91	0.143	0.172
7	96.58	95.63	0.121	0.148
8	97.02	96.12	0.103	0.131
9	97.31	96.48	0.089	0.117
10	**97.56**	**96.84**	**0.076**	**0.104**

The training and validation accuracies increase progressively across the epochs, while the corresponding losses decrease. The relatively small difference between training and validation accuracy at the final epoch suggests that severe overfitting is not evident in this illustrative experiment.

Table 5. Confusion Matrix of LSTM Classification

Assuming three classes—**Cyberbullying, Non-Cyberbullying, and Neutral/Other**—the following confusion matrix provides a consistent example:

Actual / Predicted	**Cyberbullying**	**Non-Cyberbullying**	**Neutral/ Other**	**Total**
Cyberbullying	1,922	45	33	2,000
Non-Cyberbullying	38	2,164	48	2,250
Neutral/ Other	21	30	620	671
Total	**1,981**	**2,239**	**701**	**4,921**

Note: The confusion-matrix values above correspond to an illustrative test subset and are not intended to exactly reproduce the percentages in Table 3. For a real paper, the confusion matrix should be generated directly from the actual y_test and model predictions.

Table 6. Region-Wise Prediction Results

Region	**Number of Samples**	**Cyberbullying Samples**	**Percentage (%)**
Asia	2,850	1,145	40.18
Europe	1,620	598	36.91
North America	1,940	827	42.63
South America	1,180	474	40.17
Africa	1,070	436	40.75
Oceania	840	318	37.86
Total	**9,500**	**3,798**	**39.98**

The regional distribution shows that cyberbullying-related messages occur across all regions represented in the illustrative dataset. North America has the highest proportion in this example, whereas Europe has the lowest.

International Journal of Artificial Intelligence and Communication Networks
Volume: 02, Issue: 03, July-September 2026, pp: 99-109
ISSN: 3071-3358

Table 7. Gender-Wise Distribution

Gender Category	Total Samples	Cyberbullying Samples	Percentage (%)
Male	4,620	1,895	41.02
Female	4,180	1,642	39.28
Other/Unknown	700	261	37.29
Total	**9,500**	**3,798**	**39.98**

The illustrative results show a relatively similar distribution across male and female categories. This indicates that cyberbullying-related content is not confined to a single gender category within the sample.

Table 8. Age-Wise Distribution

Age Group	Total Samples	Cyberbullying Samples	Percentage (%)
13–17	1,850	856	46.27
18–24	2,680	1,155	43.10
25–34	2,420	906	37.44
35–44	1,510	523	34.64
45+	1,040	358	34.42
Total	**9,500**	**3,798**	**39.98**

The illustrative age-wise results indicate a higher proportion of cyberbullying-related samples in the younger age categories. The 13–17 group has the highest proportion, followed by the 18–24 group.

Table 9. Ethnicity-Wise Distribution

Ethnicity Category	Total Samples	Cyberbullying Samples	Percentage (%)
Category A	2,100	852	40.57
Category B	1,950	761	39.03
Category C	1,720	695	40.41
Category D	1,480	577	38.99
Category E	1,310	522	39.85
Unknown/Other	940	391	41.60
Total	**9,500**	**3,798**	**39.98**

The illustrative ethnicity-wise distribution is relatively balanced, with no single category showing an extremely different proportion of cyberbullying-related samples. Because ethnicity is a sensitive characteristic, these results should be interpreted only as a description of the dataset and should not be generalized to broader populations.

Table 10. Comparative Model Performance

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Naïve Bayes	88.42	87.96	86.75	87.35
SVM	91.67	91.14	90.82	90.98
Random Forest	92.84	92.37	91.95	92.16
LSTM	**96.84**	**96.57**	**96.21**	**96.39**

The illustrative comparison indicates that the LSTM model outperforms the conventional machine-learning models across all four evaluation measures. Compared with Random Forest, the LSTM model provides an improvement of **4.00 percentage points in accuracy** and **4.23 percentage points in F1-score**. This improvement can be attributed to the ability of LSTM to learn sequential relationships and contextual dependencies within textual data.

Suggested graph values

For your **performance comparison graph**, use:

Model	Accuracy	Precision	Recall	F1-Score
Naïve Bayes	88.42	87.96	86.75	87.35
SVM	91.67	91.14	90.82	90.98
Random Forest	92.84	92.37	91.95	92.16
LSTM	**96.84**	**96.57**	**96.21**	**96.39**

Future Enhancements

In the future, the cyberbullying detection system can be extended to support multilingual and cross-platform analysis, enabling detection across diverse languages and social media networks. Integration of transformer-based architectures such as BERT or RoBERTa could further enhance contextual understanding and semantic accuracy. Incorporating audio and image-based bullying detection will make the system more comprehensive for multimedia content. Real-time API deployment can allow integration with live chat applications or comment sections. The use of explainable AI (XAI) can help interpret model decisions and improve transparency. Adaptive online learning mechanisms can help the system evolve with changing slang and language styles. Introducing graph-based user behavior analysis can identify repeated offenders or coordinated harassment patterns. A dashboard interface can be developed for administrators to monitor and manage detected cases. Future research may focus on reducing bias and improving fairness across demographic

groups. Overall, these enhancements will strengthen the system's scalability, accuracy, and social impact.

6. Conclusion

The proposed cyberbullying detection framework effectively combines Natural Language Processing (NLP) techniques with Long Short-Term Memory (LSTM) networks to identify harmful online behavior. By leveraging advanced preprocessing methods such as tokenization, stemming, lemmatization, and stopword removal, the system ensures clean and meaningful input data. Embedding-based feature extraction preserves contextual and semantic relationships, allowing the model to interpret language nuances more effectively. The LSTM architecture captures temporal dependencies and emotional tones, making it well-suited for text-based abuse detection. Experimental evaluation demonstrates high accuracy and improved classification performance compared to traditional machine learning methods. Addressing class imbalance through resampling further enhances fairness and reliability. The system contributes to building safer digital spaces by automating the detection of offensive and abusive content. It also provides a foundation for integrating real-time moderation tools in social media and chat platforms. The study highlights the crucial role of deep learning in understanding human language and emotions. Additionally, the framework can be extended to multilingual and cross-domain datasets for global applicability. Overall, this project showcases how intelligent text analysis can play a pivotal role in mitigating cyberbullying and promoting positive online communication.

References

[1]. Y. Hu, E. M. Clancy, and B. Klettke, "Understanding the vicious cycle: Relationships between nonconsensual sexting behaviours and cyberbullying perpetration," Sexes, vol. 4, no. 1, pp. 155–166, Feb. 2023, doi: 10.3390/sexes4010013.

[2]. M. S. Uddin and M. D. Zainlabuddin, "Secure video processing and watermark embedding using Shamir secret rule," Int. J. Artif. Intell. Comput. Electron., vol. 2, no. 1, pp. 25–31, Mar. 2026.

[3]. L. H. Collantes, Y. Martafian, S. N. Khofifah, T. K. Fajarwati, N. T. Lassela, and M. Khairunnisa (2020), "The Impact of Cyberbullying on Mental Health of the Victims," Proc. 4th Int. Conf. Vocational Educ. Training (ICOVET), pp. 30–35.

[4]. T. Ahmed, M. Kabir, S. Ivan, H. Mahmud, and K. Hasan (2021), "Am I Being Bullied on Social Media? An Ensemble Approach to Categorize Cyberbullying," Proc. IEEE Int. Conf. Big Data (Big Data), pp. 2442–2453.

[5]. Sannidhanam, A. H. (2026). LLM-Driven Workflow Optimization: Intelligent Routing in Asynchronous Distributed Systems. International Journal of AI and Machine Learning, 1(2), 23–33.

[6]. S. Joshi, H. G. Nagariya, N. Dhanotiya, and S. Jain, "Identifying fake profile in online social network: An overview and survey," Commun. Comput. Inf. Sci., vol. 1240, pp. 17–28, Jan. 2020, doi: 10.1007/978-981-15-6315-7_2.

[7]. E. Vogels, Teens and Cyberbullying 2022, Pew Research Center, Dec. 23, 2024. [Online]. Available: https://www.pewresearch.org/internet/2022/12/15/teens-and-cyberbullying-2022/

[8]. Dr. Abdul Bari, Dr. Imtiyaz Khan, Dr. Rafath Samrin, Dr. Akhil Khare, "VPC & Public Cloud Optimal Performance in Cloud Environment," Educational Administration: Theory and Practice, ISSN No: 2148-2403, Vol. 30, Issue 6, June 2024.

[9]. P. Yi and A. Zubiaga (2023), "Session-Based Cyberbullying Detection in Social Media: A Survey," Online Social Netw. Media, vol. 36, Art. no. 100250.

[10]. Anjani Haritha Sannidhanam. (2023). Self-Healing Distributed Systems: AI-Driven Failure Prediction and Automated Recovery. International Journal of Research & Technology, 11(4), 168–174.

[11]. R. Gün and G. G. Akduman, "What is cyberbullying?" in Bullying in Media and Beyond. Turkey: IGI Global, pp. 473–485, doi: 10.4018/978-1-6684-5426-8.CH028.

[12]. I. Ali and N. Hameed (2017), "Hybrid Tools and Techniques for Sentiment Analysis: A Review," Int. J. Multidisciplinary Sci. Eng., vol. 8, no. 4, pp. 28–33. [Online]. Available: https://www.researchgate.net/publication/318351105

[13]. Distributed Data Processing Frameworks for Large-Scale Healthcare Analytics. (2019). International Journal of Advance Industrial Engineering, 7(04), 1–10. doi:10.14741/ijaie/v.7.4.01.

[14]. T. Ahmed, S. Ivan, M. Kabir, H. Mahmud, and K. Hasan (2022),

"Performance Analysis of Transformer-Based Architectures and Their Ensembles to Detect Trait-Based Cyberbullying," Social Netw. Anal. Mining, vol. 12, no. 1, p. 99.

[15]. S. Unnava and S. R. Parasana (2024), "A Study of Cyberbullying Detection and Classification Techniques: A Machine Learning Approach," Eng. Technol. Appl. Sci. Res., vol. 14, no. 4, pp. 15607–15613.

[16]. Anjani Haritha Sannidhanam. (2024). Guardrails and Safety Mechanisms for LLM-Powered Enterprise Applications. International Journal of Engineering Science & Humanities, 14(3), 273–285.

[17]. M. D. Zainlabuddin and N. Sharma, "Security enhancement in data propagation for wireless network," Rev. GEINTEC-Gestão, Inovação e Tecnologias, vol. 11, no. 4, pp. 4110–4119, Aug. 2021.

[18]. S. Cook, Cyberbullying Statistics and Facts for 2024, Dec. 23, 2024. [Online]. Available: https://www.comparitech.com/internet-providers/cyberbullying-statistics

[19]. J. O. Atoum (2020), "Cyberbullying Detection Through Sentiment Analysis," Proc. Int. Conf. Comput. Sci. Comput. Intell. (CSCI), pp. 292–297.

[20]. Afsha Nishat, Dr. Mohd Abdul Bari, Dr. Guddi Singh, "Mobile Ad Hoc Network Reactive Routing Protocol to Mitigate Misbehavior Node," International Journal Of Intelligent Systems And Applications In Engineering, IJUSEA, ISSN:2147-6799, Nov. 2023.

[21]. R. Endsuy (2021), "Sentiment Analysis Between VADER and EDA for the U.S. Presidential Election 2020 on Twitter Datasets," J. Appl. Data Sci., vol. 2, no. 1, pp. 8–18.

[22]. Sannidhanam, A. H. (2020). Scalable Web Services Architecture for Healthcare Data Processing. International Journal of Recent Advances in Science and Technology, 7(01), 1–14.

[23]. M. Kousar, Dr. Sanjay Kumar, Dr. Mohammed Abdul Bari, "Design of a Decentralized Authentication and Off-Chain Data Management Protocol for VANETs Using Blockchain," Communications on Applied Nonlinear Analysis, ISSN: 1074-133X, Vol. 32 No. 2, 2025.

[24]. Performance Analysis of Wireless Sensor Networks for Smart Monitoring Applications. (2016). International Journal of Humanities and Information Technology, 1(04), 10–29. doi:10.21590/ijhit.01.04.04.

[25]. D. Javed, N. Z. Jhanjhi, and N. A. Khan (2023), "Football Analytics for Goal Prediction to Assess Player Performance," Proc. Int. Conf. Innov. Technol. Sports (RevealDNA ICITS), pp. 245–257.

[26]. Event Streaming Architectures for High-Volume Transaction Processing. (2022). International Journal of Advance Industrial Engineering, 10(01), 1–8. doi:10.14741/

[27]. L. Grunin, G. Yu, and S. S. Cohen (2021), "The Relationship Between Youth Cyberbullying Behaviors and Their Perceptions of Parental Emotional Support," Int. J. Bullying Prevention, vol. 3, no. 3, pp. 227–239.

[28]. IEEE Access, vol. 12, pp. 160822–160834. [The supplied reference is incomplete; author and article title were not provided.]

[29]. Sannidhanam, A. H. (2025). Autonomous AI Agents for Cloud Infrastructure Operations. Journal of Contemporary Science and Technology Management, 1(01), 83–100.

[30]. Anjani Haritha Sannidhanam. (2023). Zero-Downtime AI Model Updates in Real-Time Inference Systems. International Journal of Engineering Science & Humanities, 13(2), 70–78.