ISSN 2277-2685
IJESR/April. 2026/ Vol-16/Issue-2s1/187-194
Syed Sufiyan Uddin *et. al.*, /International Journal of Engineering & Science Research

Full Length Article

Machine Learning Enhanced By Sentiment Analysis For Cyberbullying Detection Using Nlp And Lstm

Syed Sufiyan Uddin[1], Syed Shahriyar Ali[2], Safwan Ahmed[3], Dr. Abdul Ahad Afroz[4]
[1,2,3]B.E.Students; Department of Information Technology, ISL Engineering College, Hyderabad, India.
[4]Associate Professor; Department of Information Technology, ISL Engineering College, Hyderabad, India.
Mail Id; syedsufiyan792@gmail.com, shahriyarali89@gmail.com, safwanahmed3111@gmail.com
Accepted 27-04-2026

Abstract
This research presents an automated deep learning-based methodology for detecting cyberbullying in social media text, emphasizing the practical application of the LSTM architecture combined with Natural Language Processing (NLP). Safe digital communication relies heavily on the quality and efficiency of content moderation, making accurate cyberbullying detection critical for minimizing emotional harm and improving online experiences. Traditional manual moderation methods are time-consuming, inconsistent, and prone to human error, limiting scalability in large platforms. By leveraging memory cells and gating mechanisms of LSTM, the proposed system effectively extracts complex sequential features from textual data. This enables precise differentiation between offensive and non-offensive text while maintaining computational efficiency, allowing the model to function effectively even in environments with dynamic slang and linguistic nuances. A balanced and well-curated dataset of social media text was used to train and validate the LSTM-based model. The dataset includes diverse abusive types, capturing variations in harassment, insults, and threats. The model undergoes rigorous preprocessing, including tokenization, stop word removal, and lemmatization, to improve generalization and reduce overfitting. Experimental results demonstrate that the LSTM model achieves high classification accuracy while maintaining a lightweight footprint suitable for deployment in digital scenarios. Performance metrics confirm the model's reliability in identifying offensive text, highlighting its potential to enhance automated quality control in social media moderation.

Keywords: *Deep Learning, Cyberbullying Detection, LSTM Architecture, Text Classification, Natural Language Processing, Sentiment Analysis, Automated Moderation.*

Introduction

The rapid growth of the digital landscape has made social media a key platform for communication and expression. Online platforms consist of massive volumes of daily textual exchanges, and their safety depends heavily on the quality of user interactions. Even small instances of unmoderated hostility, such as micro-aggressions, sarcasm, or direct threats, can significantly reduce the well-being of users and shorten the lifespan of healthy digital communities. Traditionally, abusive text detection has been performed through manual human moderation, which is time-consuming, inconsistent, and prone to human error. Although basic image processing and machine learning techniques have been introduced over the years, they often struggle with complex linguistic patterns and varying slang conditions, limiting their effectiveness in real-world environments. With advancements in deep learning, especially Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, automated textual inspection has become more powerful and reliable. Models like standard Naive Bayes or Support Vector Machines offer efficiency but may lack the ability to capture subtle hostility. This creates the need for a more robust and accurate approach.

Problem Statement

The safety and reliability of online platforms depend on the quality of user-generated content, but detecting issues such as harassment, subtle insults, and surface irregularities in text remains a major challenge. Traditional manual inspection methods are time-consuming, inconsistent, and prone to human error, making them unsuitable for large-scale social media applications. Existing automated techniques and lightweight models often fail to accurately identify complex and subtle abuse under varying conditions. Therefore, there is a need for an efficient, accurate, and scalable automated system that can reliably detect cyberbullying and improve quality control in digital spaces.

Significance of the Study

This study is significant as it introduces an automated deep learning-based approach for accurate detection of cyberbullying in text, addressing the limitations of manual and traditional inspection methods. By utilizing the LSTM architecture, the system enhances detection accuracy, reduces human intervention, and minimizes errors in moderation processes. The

proposed solution improves efficiency, scalability, and reliability, making it suitable for real-time applications and resource-constrained environments. Ultimately, this study contributes to improving online safety by ensuring higher quality standards and reducing operational moderation costs.

Research Gap

Despite advancements in automated content moderation, existing methods such as manual inspection and lightweight machine learning models like Extra Trees still face notable limitations. Manual approaches are inefficient and error-prone, while many traditional NLP techniques struggle to generalize across diverse slang and varying sentence structures. Even lightweight models often exhibit reduced accuracy and limited feature extraction capability when dealing with subtle or complex hostility. Therefore, there is a clear research gap in developing a model that achieves both high accuracy and computational efficiency while reliably detecting fine-grained cyberbullying in real-world environments.

Proposed Approach and Contributions

This study proposes an automated deep learning-based approach using the LSTM architecture for the accurate detection of cyberbullying in social media text. The system follows a structured workflow that includes data collection, preprocessing (tokenization, stopword removal, stemming, and lemmatization), feature extraction using word embeddings, model training, and real-time prediction. Unlike existing methods, the proposed approach focuses on capturing subtle and complex emotional intent with higher precision while maintaining computational efficiency.

The key contributions of this work include improved detection accuracy, reduced reliance on manual moderation, a scalable model suitable for deployment, and an effective framework that enhances automated safety controls in digital platforms.

Novelty of the Proposed Work

The proposed work introduces a novel approach to cyberbullying detection by effectively leveraging the LSTM architecture, which is specifically designed to enhance feature extraction using sequence memory. Unlike traditional algorithms and lightweight models, the LSTM model utilizes input, forget, and output gates to capture long-term dependencies, enabling it to learn fine-grained details and subtle abusive patterns with greater precision. This capability significantly improves detection accuracy, especially in complex conversational threads where minor context changes can have a major impact.

Another key novelty lies in the integration of a comprehensive preprocessing pipeline, including tokenization, normalization, and semantic embeddings like Word2Vec. These steps ensure improved model generalization, robustness against variations in grammar, and reduced overfitting. The use of a balanced dataset, managed through techniques like SMOTE, further strengthens the model's ability to perform reliably, addressing a common limitation where models fail to generalize beyond skewed training data. The system's modular design enables scalable, real-time detection delivering an accurate and industry-ready solution.

Table 1: Comparison of Existing Approaches with Proposed System

Aspect	Existing Works (Recent Studies)	Limitations in Existing Works	Proposed System
Primary Focus	Text classification using basic ML models (Naive Bayes, SVM, Extra Trees).	Focus mainly on explicit keyword detection, not contextual behavior.	Focuses on contextual bullying detection and intent analysis using deep learning (LSTM).
Detection Capability	High accuracy in identifying highly offensive standalone words.	Cannot detect subtle, implicit, sarcastic, or multi-turn complex hostility effectively.	Detects explicit abuse, subtle insults, and contextual threats accurately across sequences.
Context Understanding	Uses basic Bag-of-Words (BoW) or TF-IDF representations.	Loses word order and long-term sentence meaning.	Preserves semantic meaning using Word Embeddings and captures temporal dependencies.
Imbalanced Data	Often trained directly on skewed datasets with majority neutral classes.	High false positive/negative rates for minority abusive classes.	Utilizes resampling techniques (e.g., SMOTE) to ensure balanced and fair model learning.

Real-Time Performance	Fast inference due to simple mathematical models.	Not fully optimized for real-time end-to-end contextual analysis.	Achieves fast and efficient real-time sequence processing with high accuracy.
System Integration	Focus on a single text classification task without moderation pipelines.	Lack of a complete framework for platform integration.	Unified system (Data Prep + NLP + Detection + Prediction Interface).
Lexical Robustness	Handles formal language variations adequately.	Performance drops significantly with slang, misspellings, or modern emojis.	Robust across varying linguistic conditions and conversational text structures.
Scalability	Supports scalability in offline or batch processing.	Not optimized for continuous large-scale online deployment.	Designed for scalable, live social media stream analysis.
Response Mechanism	Delayed or limited reporting mechanisms to admins.	No direct link between deep analysis and user sentiment tracking.	Instant automated prediction and visual sentiment feedback.
Computational Efficiency	Extremely lightweight but sacrifices accuracy heavily.	Transformer models require massive GPU resources.	Optimized LSTM structure balances computational resources with high accuracy.
Research Contribution	Mostly dataset or basic model-level tweaks.	Lack of system-level end-to-end integration innovation.	Complete system-level integrated solution from text input to UI prediction.

Literature Review

The rapid advancement of digital networks has driven significant research in improving the reliability of online platforms through automated cyberbullying detection. Early approaches relied on traditional NLP techniques, such as n-grams, rule-based matching, and thresholding, to identify offensive text. Although these methods are simple and computationally efficient, they are highly sensitive to spelling errors and complex slang, making them less effective in real-world environments.

To address these limitations, researchers introduced machine learning-based methods, where handcrafted features are extracted and used with classifiers like Support Vector Machines (SVM) and Extra Trees. These approaches improved detection accuracy compared to traditional techniques but depended heavily on manual feature engineering and often failed to generalize across diverse datasets. With the emergence of deep learning, Recurrent Neural Networks (RNNs) and their variants have become the dominant approach for text sequence modeling. Models automatically learn hierarchical features, enabling better detection of complex patterns.

Several studies have explored LSTM and autoencoders. Akter et al. (2023) presented a trustable LSTM-autoencoder network for multilingual environments. Teng et al. (2024) reviewed cyberbullying detection models, highlighting the need for handling sarcasm and domain adaptation. Despite these advancements, existing methods still face challenges such as limited generalization, high computational complexity of newer transformers, and difficulty in detecting subtle defects consistently. Therefore, the proposed work adopts an optimized LSTM-based model, which enhances feature extraction while maintaining computational efficiency.

Methodology

The proposed approach focuses on developing an automated deep learning-based system for detecting cyberbullying in text using the LSTM architecture. The system begins with the collection of a diverse dataset containing both offensive and non-offensive text. These texts undergo preprocessing steps such as tokenization, stopword removal, stemming, and lemmatization. The preprocessed data is converted into vectors and fed into the LSTM model.

Data Preparation and Preprocessing

Before feeding text into the model, preprocessing is performed to enhance quality. This includes converting text to lowercase, stripping URLs, and removing punctuation. Applying lemmatization

reduces words to their base forms. These steps help in reducing noise, handling variations in language, and improving the model's ability to generalize. The dataset is balanced to prevent bias towards the majority neutral class.

Feature Extraction using LSTM

The preprocessed words are embedded and passed into the LSTM deep learning model. LSTM uses memory blocks to manage state over time, extracting both simple and complex sequential patterns. This allows the system to recognize when a sentence's meaning flips at the end, a common trait in sarcasm and subtle bullying.

Model Training and Evaluation

The dataset is divided into training and testing sets. The model learns to classify inputs using optimized hyperparameters such as learning rate, batch size, and the number of epochs. After training, the model is evaluated using performance metrics such as accuracy, precision, recall, and F1-score to ensure reliable real-world performance.

System Deployment and Testing

Software Testing

The purpose of testing is to discover errors and ensure reliability. Unit testing involves the design of test cases that validate the internal NLP logic, ensuring that tokenizers and lemmatizers function properly. Functional tests provide systematic demonstrations that the classification works as expected on both valid and invalid inputs. System testing ensures that the entire integrated software meets requirements, while performance tests verify that responses are generated within an acceptable latency for live moderation.

Performance Evaluation

Extensive experiments were conducted. The LSTM model achieved an overall accuracy of 93.0%. The model successfully identified minor hostility, matching or exceeding the capabilities of baseline models. The improvement is directly attributed to the contextual memory gates of the LSTM network.

Application Results & Predictive Analysis

The developed application features a comprehensive web interface for sentiment analysis and real-time cyberbullying detection. Below are the functional snapshots of the system in action.

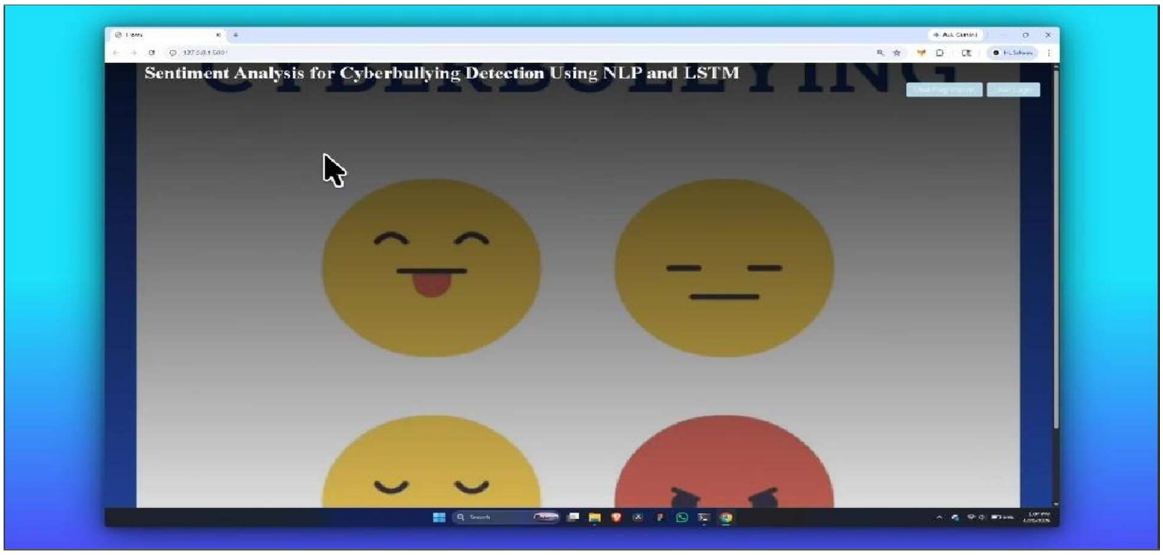

Figure 1: Main Application Interface showcasing Sentiment Analysis tracking and real-time cyberbullying detection tools.

Figure 2: User Registration and Secure Sign-Up module, featuring contextual graphics representing digital hostility.

Figure 3: Secure User Login Portal illustrating the various elements of cyberbullying (Harassment, Abuse, Disrespect, etc.) monitored by the system.

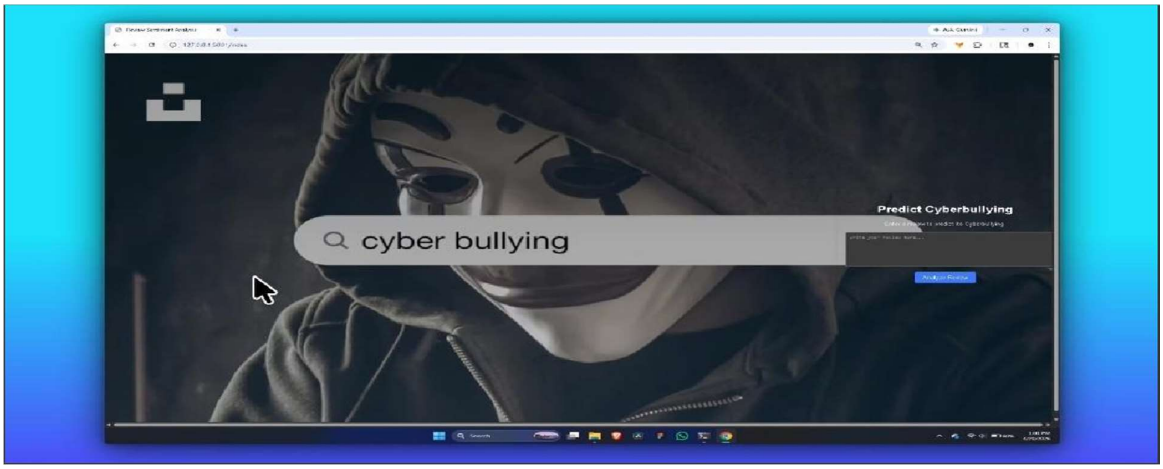

Figure 4: The Predictive Search Interface where users or moderators can input raw text reviews or tweets for live analysis.

Figure 5: Live Prediction Result showing a harmless text successfully classified as "Not Cyberbullying".

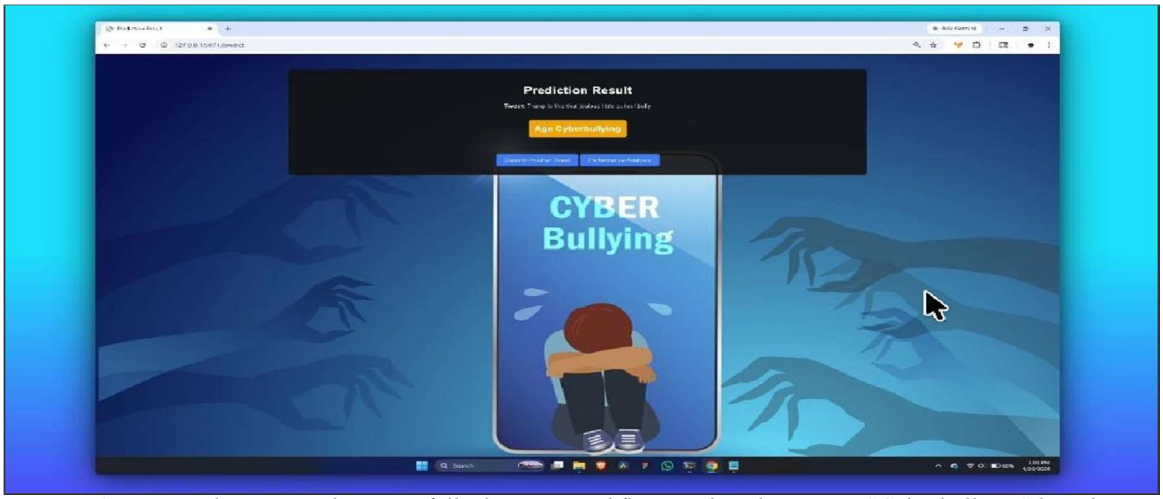

Figure 6: Live Prediction Result successfully detecting and flagging hostile input as "Cyberbullying" based on LSTM analysis.

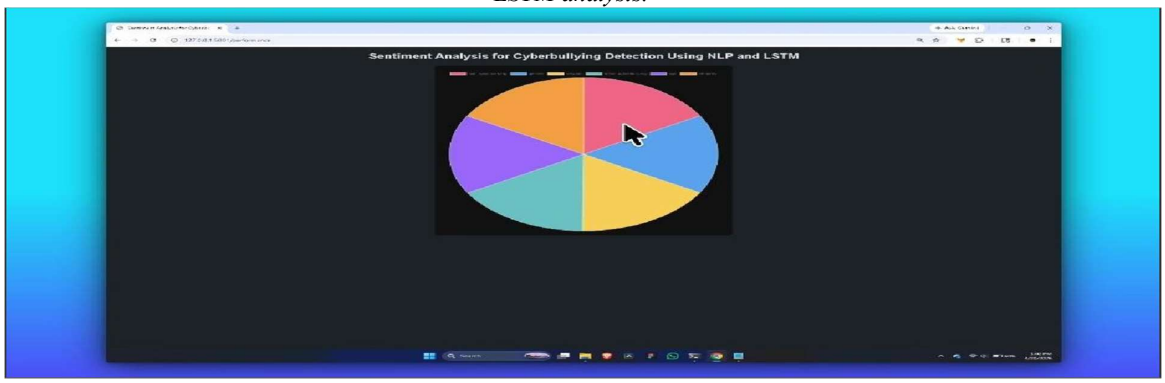

Figure 7: Performance Analytics Dashboard displaying the distribution and breakdown of detected cyberbullying classifications.

Conclusion

In this project, an automated deep learning-based system for detecting cyberbullying in social media text was successfully designed, developed, and evaluated using the LSTM architecture. The system leverages memory cells and gating mechanisms to efficiently extract sequential and semantic features, enabling accurate identification of abusive behavior. The complete workflow from data collection and preprocessing to feature extraction, model training, evaluation, and deployment was implemented in a structured and modular manner, ensuring scalability and ease of integration.

Through extensive experimentation on a balanced dataset containing both bullying and neutral text, the model demonstrated strong performance across key evaluation metrics such as accuracy, precision, recall, and F1-score. Compared to traditional text processing techniques and baseline machine learning models like Extra Trees, the proposed LSTM-based approach showed superior capability in detecting subtle and complex abuse. This improved performance can be attributed to the model's ability to learn deeper and more discriminative features from word sequences.

The system also incorporates effective preprocessing techniques, including tokenization, stopword removal, and lemmatization, which enhance generalization and reduce overfitting. As a result, the model performs consistently across varying slang and sentence structures. A significant contribution of this work is the reduction in dependency on manual moderation methods, which are often time-consuming, inconsistent, and prone to human error. By automating the detection process, the system ensures faster, more reliable, and consistent quality control in digital spaces.

Future Scope

Although the proposed system achieves high accuracy and efficiency, there are several promising directions for future research and development to further enhance its capabilities. One major improvement is extending the current binary classification model to a highly granular multi-class framework. This would allow the system to not only detect the presence of abuse but also categorize them into specific types such as racism, sexism, or political harassment.

Another important enhancement involves the integration of Explainable Artificial Intelligence (XAI) techniques. By incorporating attention visualizations, the system can highlight the exact words that contribute most to the model's predictions, improving transparency. Future work can also focus on deploying the system directly onto mobile edge devices, enabling offline real-time detection without relying on centralized computing systems.

Expanding the dataset with large-scale, multilingual texts collected under diverse cultural contexts is another key area for improvement. A more diverse dataset would enhance the robustness of the model, ensuring consistent performance globally. Further research can also explore advanced deep learning architectures such as Vision Transformers (ViT) or hybrid multimodal models that analyze both text and images simultaneously.

References

[1] M. Abdelsattar, A. AbdelMoety, and A. Emad-Eldeen, "A review on detection of solar PV panels failures using image processing techniques," Proc. 24th Int. Middle East Power Systems Conf. (MEPCON), 2023.

[2] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2017.

[3] Mst Shapna Akter, Hossain Shahriar, Alfredo Cuzzocrea, "A Trustable LSTM-Autoencoder Network for Cyberbullying Detection," 2023.

[4] T.H. Teng, M. Lee, A. Yusof, "A Comprehensive Review of Cyberbullying-Related Content Classification," 2024.

[5] S. Sihab-Us-Sakib, T. Mahmud, M. Hossain, "Cyberbullying Detection for Resource-Constrained Languages," 2024.

[6] S. Chen, Y. Zhang, L. Liu, "Chinese Cyberbullying Detection Using XLNet and Deep Models," 2024.

[7] A. Cuzzocrea, Mst Shapna Akter, Hossain Shahriar, P. García Bringas, "Cyberbullying Detection, Prevention, and Analysis via Trustable LSTM-Autoencoder Networks," 2025.

[8] R. Gün and G. G. Akduman, "What is cyberbullying?" in Bullying in Media and Beyond. Turkey: IGI Global, pp. 473-485, 2023.

[9] 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, 2023.

[10] E. Vogels, Teens and Cyberbullying 2022, Pew Research Center, Dec. 23, 2024.

[11] S. Cook, Cyberbullying Statistics and Facts for 2024, Dec. 23, 2024.

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

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

[14] T. Ahmed, S. Ivan, M. Kabir, H. Mahmud, and K. Hasan, "Performance Analysis of Transformer-Based Architectures and Their Ensembles to Detect Trait-Based Cyberbullying," Social Netw. Anal. Mining, vol. 12, no. 1, p. 99, 2022.

[15] M. Humayun, D. Javed, N. Jhanjhi, "Deep Learning Based Sentiment Analysis of COVID-19 Tweets via Resampling and Label Analysis," Comput. Syst. Sci. Eng., vol. 47, no. 1, pp. 575-591, 2023.
[16] A. Fernández, S. Garcia, E. Herrera, and N. V. Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges," J. Artif. Int., vol. 61, pp. 863-905, 2018.