International Journal of Multidisciplinary Engineering in Current Research - IJMEC

Volume 11, Issue 4s, April-2026, http://ijmec.com/, ISSN: 2456-4265

Full Length Article

Sentiment Analysis For Cyberbullying Detection Using NLP And LSTM

Syed Burhan Ahmed[1], Mohd Ateeq[2], Mohd Abbu[3], Dr. Syed Asadulla Hussaini[4]

[1,2,3]B.E.Students ;Department Of Artificial Intelligence & Data Science Engineering, ISL Engineering College, Hyderabad.

[4]Associate Professor, Department Of Artificial Intelligence & Data Science Engineering, ISL Engineering College, Hyderabad.

Mail Id; 160522747039@islec.edu.in, md4ateeq@gmail.com, abbusiddiq56@gmail.com

Accepted 24-04-2026

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, stop word 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), Deep Learning, Text Classification, Sentiment Analysis, Social Media Analytics, Text Preprocessing, Word Embeddings, Multi-Class Classification.*

INTRODUCTION:

In today's digital era, the rapid growth of social media and online communication platforms has provided individuals with new ways to connect, share, and express themselves. However, this advancement has also led to the rise of harmful behaviours such as cyberbullying, which has become a major social and psychological concern worldwide. Cyberbullying involves the deliberate use of digital platforms to harass, threaten, or demean others, often leaving long-lasting emotional and mental impacts on victims. The anonymous nature of online communication makes it even more challenging to detect and prevent such behaviour effectively. Traditional methods for cyberbullying detection, primarily based on basic machine learning algorithms, often struggle to capture the subtle linguistic nuances and contextual meanings present in online text. These methods tend to rely heavily on handcrafted features, which limit their adaptability across diverse platforms and languages. To overcome these limitations, recent advancements in Natural Language Processing (NLP) and deep learning have paved the way for more intelligent and context-aware detection systems. Among these, Long Short-Term Memory (LSTM) networks have shown remarkable performance in understanding sequential data and contextual dependencies in text. By integrating NLP techniques—such as tokenization, stemming, lemmatization, and stop word removal—with LSTM architectures, this study aims to develop a robust and efficient model for cyberbullying detection. Furthermore, embedding-based feature extraction ensures that the model preserves semantic relationships, enabling deeper insight into the emotional tone and intent of messages. To address the issue of class imbalance commonly found in social media datasets, resampling techniques are also applied, ensuring fair and balanced learning across different categories of cyberbullying. The proposed system ultimately seeks to enhance accuracy, contextual understanding, and overall reliability in detecting

https://doi.org/10.63665/IJMEC.1104s.48
ISSN: 2456-4265
IJMEC 2026

abusive and harmful content online, contributing to a safer digital environment.

LITERATURE REVIEW:

Cyberbullying detection has become an important research area due to the rapid growth of social media communication and online interactions. Researchers have explored several machine learning, deep learning, and Natural Language Processing (NLP) techniques to identify abusive and harmful content effectively.

Early cyberbullying detection systems mainly relied on traditional machine learning algorithms such as Naive Bayes, Decision Trees, Support Vector Machines (SVM), and Logistic Regression. These methods used handcrafted textual features like Bag of Words (BOW) and TF-IDF for classification. Although these techniques achieved moderate accuracy, they often failed to capture contextual meaning, sarcasm, and sequential dependencies present in cyberbullying language.

To improve text understanding, researchers introduced NLP-based preprocessing techniques including tokenization, stop word removal, stemming, and lemmatization. These methods enhanced text quality and reduced noise in datasets. Sentiment analysis was also incorporated to identify negative emotional patterns associated with cyberbullying behaviour. However, traditional NLP methods alone were insufficient for understanding complex semantic relationships in social media text.

METHODOLOGY:

The proposed system uses Natural Language Processing (NLP) and Long Short-Term Memory (LSTM) networks to detect cyberbullying from social media text. Initially, the dataset containing bullying and non-bullying comments is collected from online sources. The raw textual data is then preprocessed using NLP techniques such as lowercasing, tokenization, stop word removal, stemming, and lemmatization to remove noise and improve text quality.

After preprocessing, sentiment analysis is performed to identify the emotional polarity of the text, such as positive, negative, or neutral sentiment. The cleaned text is converted into numerical representations using word embedding techniques, which preserve semantic relationships between words.

MODULES EXPLANATION:

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 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

https://doi.org/10.63665/IJMEC.1104s.48
ISSN: 2456-4265
IJMEC 2026

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.

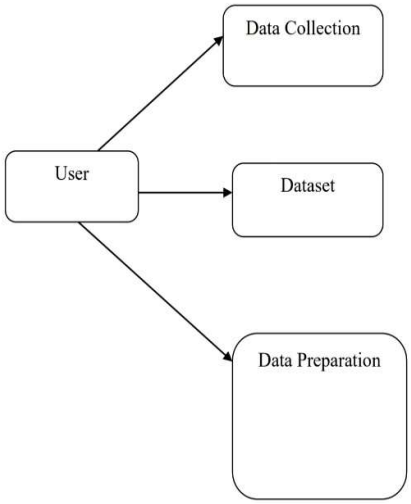
Saving the Trained Model:-

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.

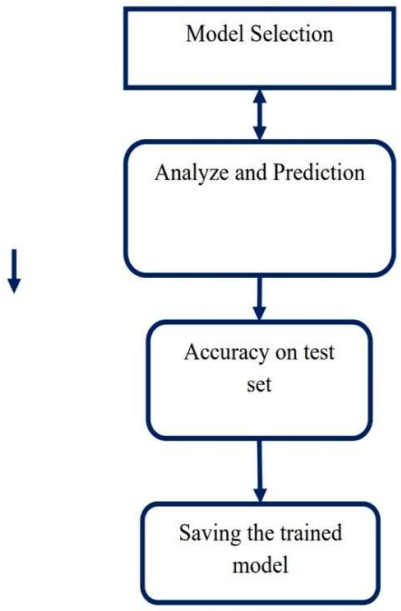
IMPLEMENTATION:

DATA FLOW DIAGRAM

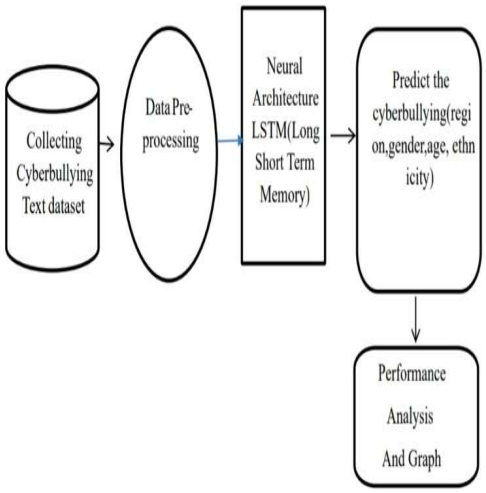
Level 0



Level 1

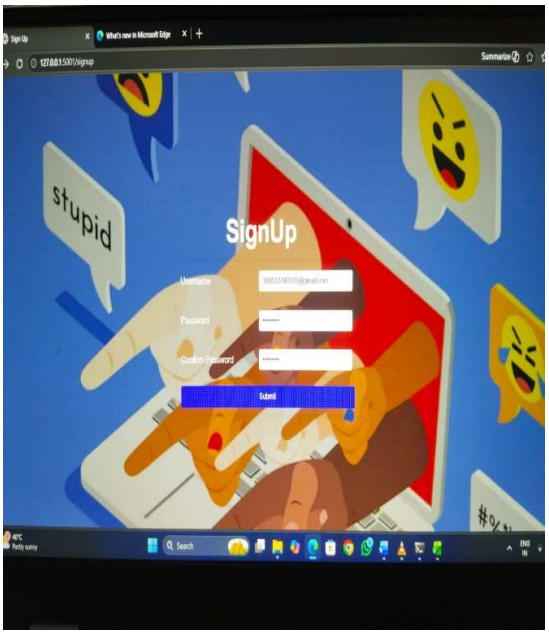


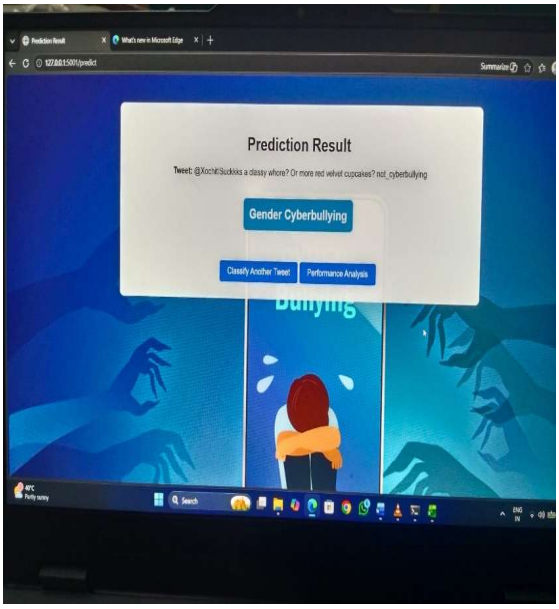
SYSTEM ARCHITECTURE:



RESULT:

https://doi.org/10.63665/IJMEC.1104s.48
ISSN: 2456-4265
IJMEC 2026





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 Robert a 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 behaviour 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.

CONCLUSION:

The proposed cyberbullying detection framework effectively combines Natural Language Processing (NLP) techniques with Long Short-Term Memory (LSTM) networks to identify harmful online behaviour. By leveraging advanced preprocessing methods such as tokenization, stemming, lemmatization, and stop word 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]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.

[2] 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.

https://doi.org/10.63665/IJMEC.1104s.48
ISSN: 2456-4265
IJMEC 2026

[3] E. I. Galyashina and V. D. Nikishin, "The concepts of aggressive information impact through the lens of internet users' worldview security," J. Siberian Federal Univ. Humanities Social Sci., vol. 14, no. Il, pp. 1660-1673, Nov. 2021, doi: 10.17516/1997-1370-0848.

[4] 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: 10.1007/978-98115-6315-7 2.

[6] E. Vogels, Teens and Cyberbullying 2022, Pew Research Center, Dec. 23, 2024. This report examines trends in teen cyberbullying and online harassment, providing statistical insights into exposure and frequency. It emphasizes the role of social media in facilitating bullying and the psychological impact on victims. [Online]. Available: https://www.pewresearch.org/internet/2022/12/15/teens-and-cyberbullying-2022/

[7] S. Cook, Cyberbullying Statistics and Facts for 2024, Dec. 23, 2024. This online article provides a comprehensive overview of cyberbullying prevalence, victim demographics, and platform-specific risks. It highlights the increasing trend of harassment on popular social media networks. [Online]. Available: https://www.comparitech.com/internet-providers/cyberbullying-statistics

[8] 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. This study investigates how cyberbullying negatively affects mental health, identifying stress, anxiety, and social withdrawal as common consequences.

[9] 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. The paper evaluates multiple ML techniques for automated cyberbullying detection, comparing SVM, Random Forest, and Naïve Bayes classifiers.

[10] 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. The study explores sentiment analysis models for detecting polarizing language, which can also inform cyberbullying detection strategies.

[11] 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. This research highlights the protective effect of parental support against cyberbullying incidents.

[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. The paper reviews hybrid sentiment analysis methods and their relevance for identifying harmful online content. [Online]. Available: https://www.researchgate.net/publication/318351105

[13] J. O. Atoum (2020), Cyberbullying Detection Through Sentiment Analysis, Proc. Int. Conf. Comput. Sci. Comput. Intell. (CSCI), pp. 292–297. The study applies sentiment analysis techniques to detect offensive messages on social media platforms.

[14] P. Yi and A. Zubiaga (2023), Session-Based Cyberbullying Detection in Social Media: A Survey, Online Social Netw. Media, vol. 36, Art. no. 100250. This survey categorizes session-based detection approaches and identifies research gaps for real-time detection systems.

[15] 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. The paper evaluates transformer models for identifying cyberbullying traits in social media text.

[16] 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. This research proposes an ensemble of ML models to improve cyberbullying detection accuracy.

[17] UNICEF (2024), Children at Increased Risk of Harm Online During Global COVID-19 Pandemic, Dec. 24, 2024. [Online]. Available: https://www.unicef.org/press-releases/children-increased-risk-harm-online-during-global-covid-19-pandemic

. The report examines rising risks of online abuse among children during pandemic lockdowns.

[18] 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. The study applies ML techniques to performance analysis, highlighting data-driven prediction strategies, which relate to sequence modeling used in cyberbullying detection.

[19] D. Javed, N. Z. Jhanjhi, and N. A. Khan (2023), Explainable Twitter Bot Detection Model for Limited Features, Proc. Int. Conf. Green Energy Comput. Intell. Technol. (GEn-CITy), pp. 476–481.

https://doi.org/10.63665/IJMEC.1104s.48
ISSN: 2456-4265
IJMEC 2026

The paper emphasizes explainable AI approaches for text classification, relevant for identifying malicious accounts in social media.

[20] D. Javed, N. Jhanjhi, N. A. Khan, S. K. Ray, A. A. Mazroa, F. Ashfaq, and S. R. Das (2024), Towards the Future of Bot Detection: A Comprehensive Taxonomical Review and Challenges on Twitter/X, Comput. Netw., vol. 254, Art. no. 110808. The study reviews bot detection, a critical step for reducing sources of cyberbullying content.

[21] M. Humayun, D. Javed, N. Jhanjhi, M. F. Almufareh, and S. N. Almuayqil (2023), 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. This paper applies deep learning with resampling techniques, a method also used to address class imbalance in cyberbullying datasets.

[22] M. F. Almufareh, N. Jhanjhi, N. A. Khan, S. N. Almuayqil, M. Humayun, and D. Javed (2024), BertSent: Transformer-Based Model for Sentiment Analysis of Penta-Class Tweet Classification, IEEE Access, vol. 12, pp. 196803–196817. The study presents a transformer-based classifier for multi-class text, improving the detection of online abusive content.

[23] F. Al-Quayed, D. Javed, N. Z. Jhanjhi, M. Humayun, and T. S. Alnusairi (2024), A Hybrid Transformer-Based Model for Optimizing Fake News Detection, IEEE Access, vol. 12, pp. 160822–160834. Although focused on fake news, the hybrid transformer methodology is applicable to cyberbullying detection for semantic understanding.

[42] A. Fernández, S. Garcia, E. Herrera, and N. V. Chawla (2018), SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-Year Anniversary, J. Artif. Int., vol. 61, pp. 863–905. The paper provides a foundation for handling imbalanced datasets, which is critical in multi-class cyberbullying detection.

[43] F. Wu, B. Gao, X. Pan, Z. Su, Y. Ji, S. Liu, and Z. Liu (2023), FACapsNet: A Fusion Capsule Network with Congruent Attention for Cyberbullying Detection, Neurocomputing, vol. 542, Art. no. 126253. This model combines attention mechanisms with capsule networks for contextual text classification.

[44] M. I. Mahmud, M. Mamun, and A. Abdelgawad (2022), A Deep Analysis of Textual Features Based Cyberbullying Detection Using Machine Learning, Proc. IEEE Global Conf. Artif. Intell. Internet Things (GCAIOT), pp. 166–170. The paper analyzes textual features and compares ML methods to enhance cyberbullying detection.

[45] S. Tambe, R. Joshi, A. Gupta, N. Kanvinde, and V. Chitre (2022), Effects of Parametric and Non-Parametric Methods on High Dimensional Sparse Matrix Representations, arXiv:2202.02894. The study addresses feature representation techniques, which are important for NLP-based cyberbullying models.

https://doi.org/10.63665/IJMEC.1104s.48
ISSN: 2456-4265
IJMEC 2026