International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 06 | Jun 2026 www.irjet.net p-ISSN: 2395-0072

A Bullying Detection Framework Using BI-LSTM and NLP

Deependra Singh[1], Bhanu Pratap Singh[2]

[1]*Research scholar, Aditya College of Technology and Science, Satna, Madhya Pradesh, India*
[2] *HOD, CSE, Aditya College of Technology and Science, Satna, Madhya Pradesh, India*

Abstract - *The rapid expansion of social media platforms has led to a significant rise in cyberbullying, posing serious psychological and social challenges. Detecting such harmful content automatically has become essential due to the massive volume of user-generated data. This paper presents a cyberbullying detection framework that integrates Natural Language Processing (NLP) techniques with deep learning models, particularly Bidirectional Long Short-Term Memory (Bi-LSTM). The proposed system follows a structured pipeline that includes data collection from social media sources, text preprocessing, feature extraction through vectorization, and semantic representation using GloVe word embeddings. Both traditional machine learning algorithms and deep learning models are implemented and evaluated. Experimental results demonstrate that deep learning models, especially Bi-LSTM, outperform conventional methods in capturing contextual dependencies and identifying bullying patterns. The model is evaluated using standard performance metrics such as accuracy, precision, recall, and F1-score, achieving improved classification performance. The proposed approach provides an efficient and scalable solution for detecting cyberbullying in online environments.*

Key Words: **Cyberbullying Detection, Natural Language Processing (NLP), Bi-LSTM, Deep Learning, Machine Learning, GloVe Embeddings**

1. INTRODUCTION

The rapid growth of social media platforms such as Twitter and YouTube has significantly transformed the way people communicate and share information. However, this widespread adoption has also led to the emergence of cyberbullying, a serious online threat that involves harassment, intimidation, or abuse through digital platforms. Cyberbullying can have severe psychological, emotional, and social consequences, particularly among adolescents and young adults [1], [2].

Traditional methods for detecting cyberbullying rely heavily on manual moderation, which is time-consuming, subjective, and not scalable for large volumes of online data. As a result, automated cyberbullying detection systems using Natural Language Processing (NLP) and Machine Learning (ML) techniques have gained significant attention in recent years [3]. These systems aim to analyze textual content and classify it into bullying or non-bullying categories efficiently.

Early approaches to cyber bullying detection primarily utilized conventional machine learning algorithms such as Naïve Bayes, Support Vector Machines (SVM), and Logistic Regression, which depend on handcrafted features and shallow representations of text [4], [5]. While these methods have shown reasonable performance, they often fail to capture contextual and semantic relationships in language, limiting their effectiveness.

Recent advancements in Deep Learning (DL), particularly Recurrent Neural Networks (RNNs) and their variants such as Long Short-Term Memory (LSTM) and Bidirectional LSTM (Bi-LSTM), have demonstrated superior performance in text classification tasks [6], [7]. These models can effectively learn sequential dependencies and contextual information in textual data, making them highly suitable for detecting nuanced and implicit forms of cyberbullying.

In addition, the use of word embeddings such as GloVe (Global Vectors for Word Representation) has further enhanced the capability of NLP models by providing dense vector representations that capture semantic similarities between words [8]. Combining word embeddings with deep learning architectures enables more accurate and robust detection systems.

In this work, we propose a cyberbullying detection framework that integrates NLP techniques, GloVe embeddings, and Bi-LSTM models to improve classification performance. The proposed system is evaluated using standard metrics such as Accuracy, Precision, Recall, and F1-Score, demonstrating its effectiveness compared to traditional machine learning approaches.

2. RELATED WORK

In recent years, cyber bullying detection has gained significant attention due to the rapid growth of social media platforms and online communication. Advanced techniques in Natural Language Processing (NLP) and Deep Learning (DL) have been widely explored to improve detection accuracy and robustness.

Recent studies indicate that deep learning models outperform traditional machine learning approaches by automatically learning semantic and contextual representations from textual data. For instance, Hasan *et al.* [9] provided a comprehensive survey highlighting the

effectiveness of deep neural networks such as LSTM, GRU, and transformer-based models in detecting cyberbullying content. Their work emphasizes the importance of contextual understanding in improving classification performance.

Transformer-based models have shown remarkable improvements in recent research. Muneer *et al.* [10] proposed a BERT-based ensemble framework for cyberbullying detection, demonstrating superior performance compared to individual classifiers. Similarly, Roy and Mali [11] explored transfer learning techniques and showed that pre-trained models significantly enhance performance, particularly in scenarios with limited labeled data.

Hybrid models combining multiple architectures have also been widely investigated. Chen *et al.* [12] introduced a hybrid model integrating XLNet with Bi-LSTM, which effectively captures both contextual and sequential information in text, leading to improved classification accuracy. In another study, Kumar *et al.* [13] proposed a multimodal deep learning framework combining textual and visual features, achieving high performance in detecting cyberbullying across different data modalities.

Handling data imbalance and improving generalization remain key challenges in cyberbullying detection. Akter *et al.* [14] proposed an LSTM-based autoencoder framework that enhances feature representation and improves detection performance on imbalanced datasets. Furthermore, Azumah *et al.* [15] addressed adversarial cyberbullying detection using deep learning techniques, highlighting the robustness of LSTM-based architectures against manipulated or disguised harmful content.

Recent advancements also include multilingual and cross-domain cyberbullying detection. Aljohani and Yafoo [16] demonstrated that combining Bi-LSTM with transformer-based embeddings significantly improves performance across different languages and datasets. Their findings suggest that hybrid deep learning approaches provide a strong balance between accuracy and computational efficiency.

Despite these advancements, challenges such as sarcasm detection, contextual ambiguity, and computational complexity of transformer models still persist. Therefore, models like Bi-LSTM remain a practical and efficient choice for real-world cyberbullying detection systems.

3. METHODOLOGY

Proposed system architecture is shown by figure below:

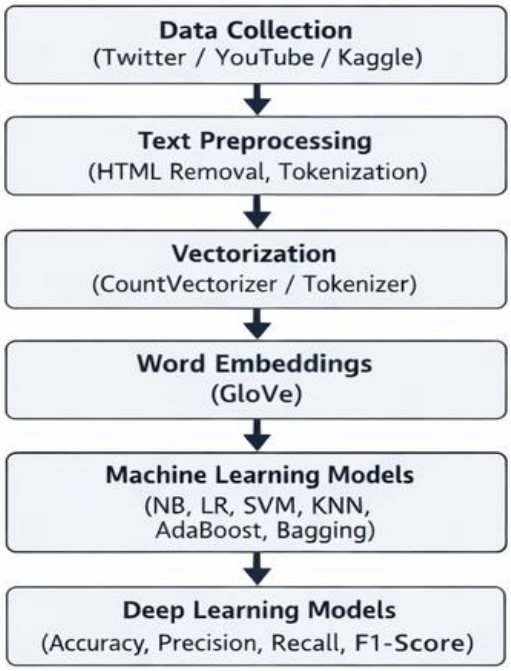


Fig-1: System architecture

The process begins with gathering textual data from platforms like Twitter, YouTube comments, and publicly available datasets (e.g., Kaggle). Raw text data is often noisy and unstructured. This stage cleans and standardizes the data by removing HTML tags, converting text to lowercase, tokenizing sentences into words, and eliminating unnecessary symbols or noise. This improves data quality and model performance. Since machine learning models require numerical input, the cleaned text is transformed into numerical representations. Techniques like CountVectorizer (word frequency-based) and Tokenizers (word-to-index mapping) are used to convert text into structured numeric form. To capture semantic meaning beyond simple word counts, GloVe embeddings are used. These convert words into dense vectors where similar words have similar representations, helping models understand context and relationships between words. Various traditional supervised algorithms such as Naive Bayes, Logistic Regression, Support Vector Machine, KNN, AdaBoost, and Bagging are applied. These models learn patterns from the vectorized data and perform classification of cyberbullying content.

Advanced deep models like LSTM and Bi-LSTM are used to capture contextual and sequential information in text. Model performance is evaluated using metrics such as Accuracy, Precision, Recall, and F1-Score, with F1-Score being particularly important for handling imbalanced datasets.

Architecture of proposed method is shown below:

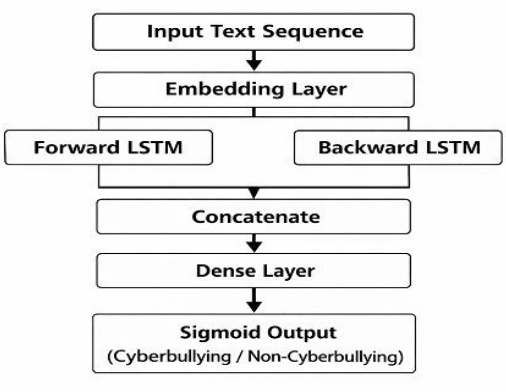


Fig-2: Proposed model

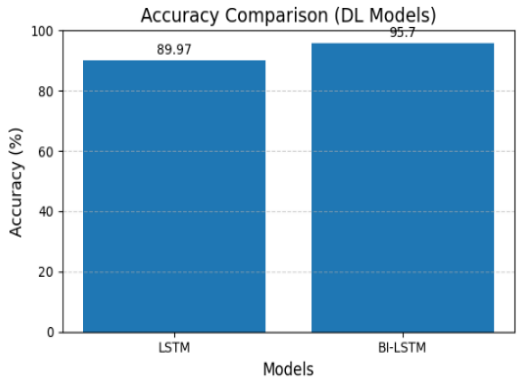
4. RESULT

After training and testing different machine learning and deep learning models, we found the values of Precision, Accuracy, F1 Score and Recall score for each model. Deep learning models (LSTM and BI-LSTM) performs better than machine learning models. Refer to Figure 5 and Figure 6 for summary of our LSTM and BI-LSTM models which we implemented. The various hyper-parameters used in the models are :

(1) Epochs- 15
(2) Dropout- 0.2
(3) Optimizer- Adam
(4) Learning Rate- 0.001
(5) Loss Function- Binary Crossentropy
(6) Batch Size- 512

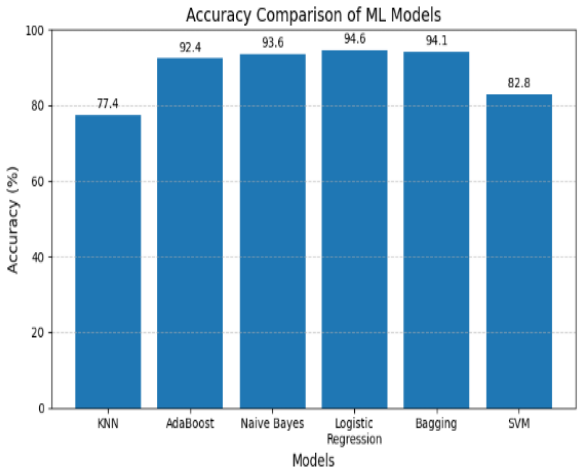
Accuracy comparison of deep models is shown below:

Chart-1: Accuracy comparison of DL models



Accuracy comparison with ML models is shown in chart below:

Chart-2: Accuracy comparison of ML models



5. CONCLUSIONS

While exploring and analyzing data we realised that it can help to weed out unnecessary information. For our dataset, we saw that keyword attributes UserIndex, Age, Annotation etc. were not present uniformly across all the datasets. As a result, we were able to remove those attributes. We also saw that the pre-processing step varies with the type of dataset. Majority of the resources that dealt with text pre-processing did not include removal of links and tags ("@"). However, the removal of these tokens was important to the dataset as they served no purpose in classifying text for cyberbullying. Deep learning models used (LSTM and BI-LSTM) with 64 neurons (>80 F1 Score) are able to learn the context better than the traditional ML algorithms. Future research can build upon this work in several meaningful directions. First, extending detection capabilities to multimodal content such as images, videos, and audio would more comprehensively address the varied nature of cyberbullying across social media.

REFERENCES

[1] K. Kowalski, R. Giumetti, A. Schroeder, and M. Lattanner, "Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth," Psychological Bulletin, vol. 140, no. 4, pp. 1073–1137, 2014.

[2] S. Hinduja and J. Patchin, "Cyberbullying: Identification, prevention, and response," Cyberbullying Research Center, 2015.

[3] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in Proc. SocialNLP, 2017, pp. 1–10.

[4] R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," in Proc. ACM SIGKDD, 2016, pp. 43–52.

[5] S. Rekha, P. K. S. Reddy, and M. S. Kumar, "Cyberbullying detection using machine learning algorithms," International Journal of Engineering and Technology, vol. 7, no. 3, pp. 1–5, 2020.

[6] Y. Kim, "Convolutional neural networks for sentence classification," in Proc. EMNLP, 2014, pp. 1746–1751.

[7] Z. Zhang, D. Robinson, and J. Tepper, "Detecting hate speech on Twitter using a convolution-GRU based deep neural network," in Proc. ESWC, 2018.

[8] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in Proc. EMNLP, 2014, pp. 1532–1543.

[9] M. T. Hasan, M. I. Hossain, and M. S. Rahman, "A review on deep learning approaches for cyberbullying detection," Future Internet, vol. 15, no. 5, pp. 1–25, 2023.

[10] A. Muneer, S. M. Anwar, and M. A. Khan, "Cyberbullying detection on social media using stacking ensemble learning and BERT," Information, vol. 14, no. 8, pp. 1–18, 2023.

[11] P. K. Roy and F. U. Mali, "Cyberbullying detection using deep transfer learning," Complex & Intelligent Systems, vol. 8, pp. 5449–5467, 2022.

[12] S. Chen, J. Wang, and K. He, "Cyberbullying detection using XLNet and Bi-LSTM hybrid model," Information, vol. 15, no. 2, pp. 1–15, 2024.

[13] A. Kumar, R. Singh, and P. Sharma, "A multimodal deep learning approach for cyberbullying detection," Applied Sciences, vol. 14, no. 24, pp. 1–20, 2024.

[14] M. S. Akter, M. Rahman, and S. Azad, "An LSTM-autoencoder framework for cyberbullying detection," IEEE Access, vol. 11, pp. 102345–102358, 2023.

[15] S. W. Azumah, K. O. Boateng, and E. K. Arthur, "Adversarial cyberbullying detection using deep learning techniques," IEEE Access, vol. 12, pp. 55678–55690, 2024.

[16] E. J. Aljohani and W. Yafoo, "Enhanced cyberbullying detection using LSTM, Bi-LSTM, and transformer-based models," IEEE Access, vol. 13, pp. 11234–11248, 2025.