RESEARCH PAPER

DEEP LEARNING BASED CYBERBULLYING DETECTION ON SOCIAL MEDIA PLATFORMS

Aeku Divyan Preethi[1], Dr. L Jagadeesh Naik[2]
[1]Department of ECE, Holy Mary Institute of Technology and Science, Hyderabad, India.
Email: divyaa.2458@gmail.com
[2]Associate Professor, Department of ECE, Holy Mary Institute of Technology and Science, Hyderabad, India.
Email: l.jagadeeshnaik@gmail.com (Corresponding Author)
***Corresponding author: Dr. L Jagadeesh Naik, Associate Professor, Department of ECE, Holy Mary Institute of Technology and Science, Hyderabad, India**
Email: l.jagadeeshnaik@gmail.com
Received: 10th April, 2026; **Revised:** 22nd April, 2026; **Accepted:** 28th April, 2026; **Available Online:** 11th June, 2026

ABSTRACT
Cyberbullying has become a major problem in social media, which poses severe psychological and emotional dangers to the users, especially adolescents and young adults. The explosive development of online communication has resulted in a rise in abusive and harmful content, which makes manual monitoring inefficient and impractical. The current systems of cyberbullying detection, which are mainly founded on the traditional machine learning techniques, are highly reliant on the detecting systems based on the analysis of keywords and lack the capacity to detect the contextual meaning, sarcasm, and linguistic variations, which contributes to the decreased accuracy and reliability. Recent works emphasize that deep learning methods are more successful than traditional ones because they are capable of capturing semantic and contextual relationships in textual data.
The proposed research involves a deep learning-based system to detect cyberbullying based on a Bidirectional Long Short-Term Memory (BiLSTM) model. The system uses Natural Language Processing (NLP) methods that include text cleaning, text tokenization, removal of stop-words, and text normalization to preprocess text in social media. The word embeddings are used to convert the textual data to meaningful numerical representations so that the model can learn the semantic relationships between words. The BiLSTM architecture handles textual data both forward and backward, enabling it to effectively deal with contextual information and sequential dependencies.
The proposed model is a multiclassification to discover the various categories of cyberbullying, in this case, age-based, gender-based, religious, and ethnic abuse. Experimental data has shown that the BiLSTM model can achieve the highest accuracy, precision, recall and F1-score than the traditional machine learning models. The system is scalable and can be deployed in real-time, which can help create safer online spaces as it can help detect and prevent cyberbullying content on the internet effectively.
Keywords: Cyberbullying detection, Deep learning, BiLSTM, Natural Language Processing, Social media analysis, Text classification.
How to cite this article: Preethi AD, Naik LJ. Deep Learning Based Cyberbullying Detection on Social Media Platforms. Int J Drug Deliv Technol. 2026;16(58s):1641-1650. DOI: 10.25258/ijddt.16.58s.174
Source of support: Nil.
Conflict of interest: None

1. INTRODUCTION

1.1 Background of Cyberbullying

Cyberbullying is defined as the utilization of digital platforms like social media, messaging apps, and online forums, to harass, threaten, or humiliate people. As the internet accessibility and social networking sites have rapidly increased, cyberbullying has become a burning global problem. Cyber bullying can happen at any time and can reach a large audience in an instant, being more harmful and persistent. Recent research points out that online harassment has grown manifold with the proliferation of platforms such as Twitter, Instagram, and Facebook where users communicate with each other via the use of text (Fati et al., 2023). The anonymity and the convenience of communication offered by these sites tend to provoke users to commit some harmful act which would not be directly punished.

1.2 Growth of Social Media and Online Communication

The rapid increase in the number of social media users has altered how individuals communicate, share information and opinions. This growth has however increased the abusive and offensive content online. The social media websites produce tremendous user-generated data per second, which can hardly be monitored manually to detect harmful content. As recent studies indicate, the magnitude and heterogeneity of interaction in online platforms have posed a huge challenge to the detection and management of cyberbullying in the most effective way possible (Mazhar et al., 2026). Moreover, the informality of online language, such as slang, abbreviations, and emojis, makes it difficult to identify the abusive content, which is a task that can be solved only with the help of sophisticated computational methods.

1.3 Psychological and Social Impact

The psychological and emotional effects of cyberbullying are severe, and as a result, the victims develop stress, anxiety, depression, and in the worst case, tend to be suicidal. This is because the impact of online harassment is heightened by its persistent nature since a victim may be exposed to harmful content over and over. Research highlights that cyberbullying has an impact on individuals of all age groups, with a special focus on adolescents and young adults, who are more active on the social media platform (Akter et al., 2023). Long-term impacts of these experiences may include low self-esteem, shun social interactions, and locate problems with school or workplace. As such, the issue of cyberbullying is not merely a technological problem, but also a social and ethical issue.

1.4 Need for Automated Detection Systems

Considering the large volume of data created on social media, manual moderation is not only inefficient and impractical, but also extremely expensive. The automated cyberbullying detection systems are now needed to detect harmful material in real time. Conventional approaches utilize human moderators or systems that are regulated by rules, which are commonly slow and inconsistent. Recent innovations in artificial intelligence and natural language processing (NLP) have given rise to the creation of automated systems that can be used to analyse large amounts of text data in an efficient manner (Gowthami and Rajesh, 2024). Such systems have the potential to help filter abusive content, lessen the work load on human moderators, and make the platforms safer.

1.5 Problems with Cyberbullying Detection

Cyberbullying is a complicated phenomenon to detect because of numerous linguistic and contextual difficulties. The offensive language is usually spoken indirectly, with the use of sarcasm, irony or coded language and therefore, it is not easily identified by the traditional systems. Also, the meaning of words may differ, according to the context, culture and the purpose of the user. Recent literature has emphasized that most of the current models have been unable to capture these contextual nuances and thus, have failed to give accurate predictions (Mohiuddin et al., 2025). Moreover, cyberbullying may be conducted in various forms, such as age, gender, religion, and ethnicity, which means that models should be able to effectively deal with multiclassification.

1.6 Deep Learning in Cyberbullying Detection

Deep learning algorithms have proven to be very promising in terms of overcoming the shortcomings of conventional machine learning algorithms. Models like Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), as well as a Bidirectional Long Short-Term Memory (BiLSTM), can comprehend sequential and contextual connections within text data. These models process text in such a way that word order is maintained and dependencies between words are captured that allow a better interpretation of meaning. Recent research shows that approaches, which are based on deep learning, are more effective in detecting cyberbullying, especially in complex and multilingual environments (Sakib et al., 2024).

1.7 Study Objective

The objective of the research is to design a deep learning based system to detect cyberbullying in social media text through a Bidirectional Long Short-Term Memory (BiLSTM) model. The suggested system is aimed at enhancing the detection accuracy, introducing NLP preprocessing methods, and text analysis within a specific context. It also seeks to categorize cyberbullying into various groups which include age, gender, religion and ethnicity. This study adds to the safer online spaces by addressing the constraints of the current systems and offering an intelligent and automated content moderation approach.

2. LITERATURE REVIEW

2.1 Introduction to the Cyberbullying Detection Research

Jaradat (2025) investigated that cyberbullying has become a sophisticated digital phenomenon impacting millions of users on the Internet every day. The paper highlighted that initial detection systems were mainly rule-based, but over time shifted to machine learning-based approaches as the amount of data grew. It further cited the fact that the detection of cyberbullying must involve an appreciation of the linguistic nuances and contextual meaning which has led researchers to consider more sophisticated and advanced computational models (Jaradat, 2025).

A systematic literature review of studies in the period 2020-2025 conducted by Fitro (2025) has revealed that the field of automated cyberbullying detection is experiencing significant improvement in a variety of modalities, including text, images, and videos. The paper emphasized the fact that the majority of the recent studies are concentrated on the text-based detection because of the prevalence of textual communication in the social media. It also indicated that cyberbullying is closely related to such psychological disorders as depression and stress, which once again suggests the significance of proper detection systems (Fitro et al., 2025).

2.2 Conventional Machine Learning Methods

Perera (2024) discussed how traditional machine learning models like Naive Bayes, Support Vector Machine (SVM), and Decision Trees could be used to detect cyberbullying. The analysis showed that these models are very dependent on structured feature extraction methods and they work reasonably well on simple datasets. But their performance is weakened in the context of complex or ambiguous text because they are unable to capture

more profound relationships that exist among semantics (Perera, 2024).

One study by Susmitha (2024) compared several machine learning models and discovered that certain algorithms such as the Random Forest and SVM generated moderate accuracy in the detection of explicit abusive language. The study stressed the fact that these methods rely on manually crafted features and need a lot of preprocessing. Although they are computationally efficient, they have difficulties with making generalizations across datasets and languages (Susmitha et al., 2024).

2.3 Feature Extraction Techniques

Ramadan (2025) dedicated attention to how the techniques of Natural Language Processing like TF-IDF and Bag of Words can be used in detecting cyberbullying. The analysis described in the research that translates textual information into numerical values based on frequency and significance of words. These methods are incapable of capturing the sentence structure and context, which limits their capability to analyse complex sentences in an accurate manner (Ramadan et al., 2025).

Unnava (2024) examined the strategies of feature engineering and emphasized that preprocessing methods have a significant impact on the performance of the models. The researchers concluded that, to a certain degree, the accuracy of detection can be enhanced by adding sentiment analysis and linguistic feature. Nevertheless, the existing traditional feature extraction methods are still not able to extract contextual meaning and semantic relationships among words (Unnava, 2024).

2.4 Limitations of traditional models

Haque (2025) has discussed the drawbacks of the traditional machine learning models in detecting cyberbullying with the emphasis that they fail to comprehend the context and sarcasm. The research observed that these models view text as separate words without taking into consideration the sentence structure and meaning. Consequently, they tend to falsely categorize minor or implicit types of cyberbullying, which results in the decreased accuracy (Hoque et al., 2025).

As Mohiuddin (2025) pointed out, the traditional systems especially fail when dealing with culturally diverse and multilingual data. The study indicated that the expressions of cyberbullying differ widely across cultures and languages, and it is hard to assume that the expressions can be identified according to the standard models. This drawback highlights the necessity of more adaptive and context-sensitive solutions (Mohiuddin et al., 2025).

2.5 Development of Deep Learning Methods

The review of the deep learning-based methods of detecting cyberbullying conducted by Hasan (2023) has shown that neural networks are superior to traditional algorithms because of their capacity to learn complex patterns automatically. The paper has highlighted that deep learning models do not require manual feature engineering but are capable of processing large-scale datasets in an efficient manner. It also found out different architectures like CNN, RNN, and LSTM to be useful in handling text classification tasks (Hasan et al., 2023).

Geetha (2025) suggested a hybrid deep learning model that combines Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks to enhance the accuracy of detection. The research revealed that a model that combines multiple architectures is more effective in modelling both the local and sequential characteristics of text data. This was an effective way to perform better than standalone models (Geetha, 2025).

2.6 Deep Learning Text Classification Models

Jaradat (2025) also discussed deep learning architectures that are specifically designed to detect cyberbullying and demonstrated that models like LSTM and BiLSTM are highly effective at capturing contextual relationship within text. The paper has emphasized the role of bidirectional processing in allowing the model to process both the past and future context, which results in a better understanding of the meaning of sentences (Jaradat, 2025).

Cuzzocrea (2025) proposed an enhanced LSTM-autoencoder model of cyberbullying detection and overcomes the problem of data inadequacy by synthesizing training data. The paper demonstrated that deep learning models can be trained to work across multiple languages and enhance their classification capabilities even in low-resource conditions, making them applicable to real-life applications (Cuzzocrea et al., 2025).

2.7 Comparative Analysis of ML and DL Approaches

Purkayastha (2025) presented a comparative study of the traditional machine learning and deep learning models and found that deep learning models are always more accurate and scalable than the traditional machine learning models. The study emphasized that deep learning models have the capacity to capture contextual and semantic information, which is crucial in identifying complex patterns of cyberbullying (Purkayastha et al., 2025).

Sakib (2024) discussed transformer-based models and showed that the advanced models of deep learning perform better than both conventional machine learning and previous neural network models. The paper has highlighted that contextual embeddings and attention mechanisms are very important in enhancing accuracy in detection especially in large dataset (Sakib et al., 2024).

2.8 Research Gaps and Improvements Needed

Fitro (2025) has listed several research gaps, such as the unavailability of models that can handle multimodal data and the scarcity of labelled dataset.

Another finding of the study was that most of the existing systems were binary based, and hence could not be applied in multiclass categorization, which limited their applicability in the real world (Fitro et al., 2025).

Mohiuddin (2025) highlighted the importance of culturally sensitive and context-aware models that are capable of detecting subtle and implicit kinds of cyberbullying. The study proposed that the future systems would be built with enhanced deep learning algorithms that would understand the linguistic differences and the changing patterns of communication (Mohiuddin et al., 2025).

3. PROBLEM STATEMENT

3.1 Increasing Incidents of Cyberbullying on Social Media

With the high growth rate of social media networks like Twitter, Instagram, and Facebook, cyberbullying has become more and more widespread. The simplicity of communication and anonymity that these sites offer are some of the reasons why people are encouraged to perpetrate abusive acts without responsibility. Recent research shows that the number of cases of cyberbullying is constantly increasing because of the increased use of digital means of communication, especially among adolescents and young adults (Jaradat, 2025). The sheer amount of user-generated content ensures that it is almost impossible to keep an eye on the harmful interactions in order to control the situation.

3.2 Psychological Effect on the users

The effects of cyberbullying are profound and far-reaching and they may not only affect victims psychologically and emotionally but also in other ways. The victims usually develop stress, anxiety, depression and in rare cases, develop suicidal tendencies. Prolonged exposure to harmful content may result in long-term mental health problems and withdrawal. Studies point out that cyberbullying has a significant implication on emotional well-being and can interfere with personal, academic, and professional life (Hithnawi, 2025). Thus, the need to combat cyberbullying is not merely technical but a social obligation as well.

3.3 Limitations of the Current Automated Systems

The current cyberbullying detection methods, which are mostly founded on the conventional machine learning methods, are limited in a number of ways. These systems are quite dependent on the analysis by key-words and do not focus on the true meaning of text. Consequently, they find it difficult to determine intricate patterns and contextual relations in sentences. Recent studies also highlight the fact that traditional models tend to yield inaccurate results when using subtle or implicit abusive language (Hoque et al., 2025).

3.4 Major Problems in Detection

A big problem in cyberbullying detection is the interpretation of the context of the use of words. The offensive content is often conveyed indirectly via sarcasm, irony, or coded language, and it is hard to be read by automated systems. Moreover, the differences in language, such as the use of slang, abbreviations, and expressions in more than one language, make it even more difficult to detect. Research indicates that current systems are not able to identify such variations precisely resulting in misclassification (Mohiuddin et al., 2025).

3.5 Need for an Intelligent Detection System

Considering these issues, there is a great necessity in an effective and intelligent system that would be able to automatically detect cyberbullying with the highest degree of accuracy. Through such a system, one should be in a position to read between the lines and pick the nuances of the language; the system should also be capable of adapting to the changing patterns of communication. State-of-the-art deep learning methods provide promising solutions to address these drawbacks and enhance the detection performance.

4. OBJECTIVES

4.1 Design of Automated Detection System

The main aim of this study is to come up with an automated system that can be able to detect cyberbullying in text messages using social media with a high level of efficiency and accuracy. As the number of online interactions grows, it is no longer possible to perform manual moderation, which requires intelligent mechanisms capable of processing large volumes of online interactions in real-time. An automated detection system is important to detect harmful content and provide better online experiences (Anuharini et al., 2025).

4.2 NLP Techniques

The other important objective is to use Natural Language Processing (NLP) methods in preprocessing textual data. This involves cleaning up texts, tokenization, elimination of stop-words, and normalization. These preprocessing phases are necessary in transforming raw text into a structured format to be analysed and used to train a model. The quality of the input data and the performance of the model are improved with the help of effective NLP techniques (Ramadan et al., 2025).

4.3 Implementation of BiLSTM Model

The proposed study will employ a Bidirectional Long Short-Term Memory (BiLSTM) model to detect cyberbullying. The biLSTM models can capture the contextual relationships in the text by processing information both forward and backward. This lets the system know the sequence and meaning of words in a better manner than the traditional models (Sree & Joseph, 2026).

4.4 Multiclass Classification of Cyberbullying

The proposed system will be structured to categorize cyberbullying into various groups of: age, gender, religion, and ethnicity. Such multiclassification method enables obtaining a more specific insight

into the various forms of abusive behaviour and enhances the general efficiency of the system.

4.5 Performance Improvement and Evaluation

Lastly, the study will enhance the accuracy of detection than other conventional machine learning techniques. To test reliability and efficacy in real world applications, the system will be tested using standard performance measures like accuracy, precision, recall and F1-score.

5. EXISTING SYSTEM

5.1 Conventional Machine Learning Models to detect cyberbullying

The current cyber bullying detection systems are predominantly based on the conventional machine learning algorithms, namely Naive Bayes, Support Vector Machine (SVM) and Decision Trees. The popularity of these models is because of their simpleness, efficiency in calculating, and their capability to work with structured data. Naive Bayes is a probabilistic classifier that presupposes independence of features and is especially effective in those text classification tasks that involve large volumes of data. Support Vector Machine, conversely, is a discrimination model, which determines the best hyperplanes to classify the data, and thus it is applicable in high-dimensional data like textual features. Decision Trees are hierarchical and are easy to interpret because they are made based on the values of features and have a hierarchical structure. These models are highly engineered in terms of features and preprocessing to attain good performance, despite its usefulness (Abdullah et al., 2025).

5.2 Feature Extraction Methods: TF-IDF and Bag of Words

The most common feature extraction methods used in traditional systems to transform textual data into numerical data include Term Frequency-Inverse Document Frequency (TF-IDF) and Bag of Words (BoW). TF-IDF uses weights to represent the importance of words in a document relative to a corpus whereas BoW represents the text as a collection of frequencies of words without considering their order. The methods allow machine learning models to handle textual information efficiently. Nevertheless, they do not recognize text as a set of tokens and cannot reflect contextual relations between text and words. Research has indicated that such approaches are only effective in detecting explicit abusive language but are ineffective in addressing more subtle or implicit forms of cyberbullying (Kumar et al., 2025).

5.3 Working Mechanism and Limitations of Traditional Systems

Data collection, preprocessing, feature extraction, model training, and classification are the general steps of the workflow of existing systems. Firstly, text data is cleaned and converted into numerical vectors with methods such as TF-IDF or BoW. They are then used to train machine learning models that will classify input text into categories such as abusive or non-abusive. The frequency of keywords and predetermined patterns form the major part of the classification process, which is why such systems highly rely on explicit signals of cyberbullying. Consequently, they do not have semantic knowledge and they do not have the ability to interpret context, sarcasm and indirect statements. This weakness diminishes their usefulness in real-life situations where language is dynamic and complex (IJRPR, 2025).

6. LIMITATIONS OF EXISTING SYSTEM

6.1 Lack of Contextual Understanding and Linguistic Limitations

The fact that they are not able to comprehend the context and sarcasm is one of the greatest limitations of the current cyberbullying detection systems. The classical machine learning models process text by considering individual words or tokens, but not the surrounding context in which they are used. This is a weakness because such systems cannot easily interpret sentences with an element of irony, sarcasm, or hidden meanings. Considering a sentence that might seem harmless on the basis of the single words but may have a harmful effect when interpreted in context. Studies have shown that such cases usually fall under the misclassified category due to lack of semantic understanding (Mohiuddin et al., 2025). Moreover, they do not recognize the sequence of words and the grammatical composition, since techniques like Bag of Words considers text to be disordered data. It leads to wrong semantic interpretation of such sentences in which sentence meaning is determined by word sequence.

6.2 Problems with performance and data-related problems

Another poor performance observed in the traditional systems is that they do not perform well whenever dealing with a complex or a long sentence that contains multiple expressions. Such models have problems in modeling relationships between words in longer sentences and result in lower classification accuracy. Data imbalance is another major problem because one of the classes (e.g., non-bullying) overtakes the data. This disproportion makes the model skewed towards the majority group which makes the prediction of the model unreliable. Moreover, the current systems cannot identify indirect or subtle cyberbullying because they highly depend on explicit key words and preset patterns. This makes them incapable of detecting any hidden or implicit types of abusive behaviours (Hoque et al., 2025).

6.3 Scalability and Real-World Limitations

Besides language and performance problems, conventional systems have problems of scale-ability and adaptability. These models do not lend themselves well to processing large scale and real time information that is generated on the social

media platforms. They need to go through regular retraining to conform to the changing language patterns, lingo, and culture. In addition, they are not as able to generalize to different datasets and domains because of their reliance on manual feature engineering. This leads to a substantial decrease in their effectiveness in real-life applications, which is why more advanced and adaptive approaches are required (IJEDR, 2026).

7. PROPOSED SYSTEM

7.1 Intro to Deep Learning-Based Approach

The new system presents a solution using deep learning to identify cyberbullying in text communication on social media. In contrast to the classical machine learning models, it uses the capabilities of the latest neural network architectures to learn the complex patterns and contextual relationships in textual data. Deep learning models can automatically learn features of raw data, without needing manual feature engineering. This methodology greatly enhances the capability to identify explicit and implicit types of cyberbullying. The recent research shows that the methods of deep learning are more accurate and robust than traditional models, especially when working with large and various datasets (Mohiuddin et al., 2025).

7.2 Natural Language Processing and Data Preprocessing

The system starts by preprocessing the input text by applying the techniques of Natural Language Processing (NLP). This involves cleaning of the text whereby irrelevant characters, punctuation and noise are eliminated to enhance the quality of the data. It is followed by the process of tokenization to divide the text into smaller units (e.g., words or tokens) to analyse it effectively. Stop-word removal is used to remove commonly-used words which do not add meaning to the text, and normalization standardizes the text by converting it to a standard format, such as lowercase. These preprocessing are required to make sure that the input data is organized and can be further processed by the model. Also, word embeddings can be used to transform text into dense numerical vectors that reflect semantic relationships between words, enabling the model to learn meaning instead of just frequency (Kumar et al., 2025).

7.3 BiLSTM Architecture and Multiclass Classification

The Bidirectional Long Short-Term Memory (BiLSTM) model is the main part of the proposed system. BiLSTM is a sophisticated form of Recurrent Neural Network (RNN) which operates on data in both forward and backward direction allowing it to extract context of both past and future words within a sequence. This two-way processing improves the capacity of the model to comprehend the structure of sentences and the meaning of the context. The model successfully addresses sequential dependencies, which is why it can be used to analyse complicated and long sentences. The system is a multiclass classifier, classifying text into various types of cyberbullying like age, gender, religion and ethnicity and non-cyberbullying content. This method offers a more in-depth insight into the abusive behaviours than the binary classification models.

7.4 System Workflow and Integration

The general workflow of the suggested system includes the data collection, preprocessing, representation of the features with the help of word embeddings, training of the model, and classification of the data. In training, the BiLSTM model is trained on labelled data and the model parameter is adjusted to reduce errors in prediction. After being trained, the model is capable of real-time classification of new input text. The system can be incorporated into the social media to trace the user-generated contents and automatically identify cyberbullying. This integration allows real-time intervention, including flagging or blocking harmful content, thus improving the safety of users. Deep learning models are scalable and flexible and can be used to implement in large-scale applications (Habiba et al., 2025).

8. BENEFITS OF PROPOSED SYSTEM

8.1 Better Accuracy and Context Awareness

The proposed system, based on deep learning, has much better accuracy than the traditional machine learning methods. Using models like BiLSTM, the system is able to capture the contextual relationships and comprehend the order of words in a sentence. This allows it to infer the meaning in a better way leading to a better classification performance. The proposed system can analyse semantic relationships in contrast to the traditional methods, where the frequency of keywords is used to identify cyberbullying (Mohiuddin et al., 2025).

8.2 Complex Language and Subtle Cyberbullying

The other significant benefit of the proposed system is that it can deal with complex and long sentences. The BiLSTM model is effective in processing sequential data, as it is able to determine trends in more than a single word and phrase. This will allow the system to identify subtle and implicit cyberbullying that might not be using explicit offensive language. It is also able to address language differences such as slang, shortening, and use of informal language, which is prevalent in social media (Habiba et al., 2025).

8.3 Scalability and Real-Time Application.

The proposed system is very scalable and can be used to run in real-time on a large social media platform. Deep learning models can effectively process large volumes of data and are therefore suitable to the task of tracking continuous streams of user-generated content. Also, the system minimizes bias due to the learning process based on various datasets and the capacity to adapt to various communication styles. The capability of generalizing in different contexts and languages

increases its applicability in real-life situations. In general, the suggested solution is a solid and trustworthy option to detect cyberbullying and make the Internet safer (IJEDR, 2026).

9. METHODOLOGY

9.1 Data Set and Data Preprocessing Methods

The data set in this paper is a set of labelled social media text that is gathered on social media sites like Twitter and online forums. The dataset consists of various types of cyberbullying, including age-based, gender-based, religious, and ethnic harassment. The datasets that are publicly available are the most common databases that are used in the study of cyberbullying. The pre-labelling of these datasets allows supervised learning methods to be used in classification tasks. It has been found that the quality and diversity of data sets play a crucial role in the performance of detection models (Philipo et al., 2024).

Data preprocessing is an important step in the process of preparing raw text to analysis. Starting with the text cleaning process, in which the process eliminates unwanted symbols, punctuation, and noises. This is followed by tokenization where sentences are broken down to individual words or tokens. Removal of stop words removes words that occur frequently but have no meaningful value, e.g. then, is. Normalization methods, such as lowercasing and stemming, are used to guarantee uniformity in representing the text. Preprocessing techniques make the model training more efficient and enhance the overall performance (Mazhar et al., 2026).

9.2 Feature Representation and Model Architecture

Upon preprocessing, textual content is converted into numbers with word embeddings, including Word2Vec and GloVe. The techniques of embedding embed semantic links between words by projecting them into dense vectors. In contrast to the classical approach to feature extraction, embeddings do not lose the contextual meaning, and due to this feature, the model can better comprehend the relationships among the words (IRJMS, 2024).

The main model adopted in this study is Bidirectional Long Short-Term Memory (BiLSTM) network. BiLSTM is a state with two layers of LSTMs that can process the input sequences forward and backward to allow the model to capture both the past and future context. The architecture consists of embedding layer, BiLSTM layers, dropout regularization layers, and dense output layer with a SoftMax activation function to perform multiclass classification. This construction enables the model to deal with sequential dependencies and intricate linguistic patterns efficiently (IJRASET, 2025).

9.3 Training and Evaluation Process

Supervised learning is the process of training the model based on labelled information. The training process consists of breaking up the dataset into training, validation, and testing sets, usually in an 80:10:10 ratio. The loss function used is categorical cross-entropy, which measures the difference between predicted and actual labels. Adam optimizer is used to optimize the model weights. Training is done through several epochs with a specified batch size to guarantee convergence and stability.

The model is assessed based on the common scores, including accuracy, precision, recall, and F1-score. The proportion of correctly classified instances is an accuracy measure, and the proportion of positive predictions that are correct is a precision measure. Recall is a measure of how well the model can identify all the relevant cases, and F1-score is a balanced measure of precision and recall. These measures are common across research on cyberbullying detection models to evaluate model performance in all its facets (Mazhar et al., 2026; IJRASET, 2025).

The system is implemented in Python programming language with the use of libraries like TensorFlow, Keras and Scikit-learn. These tools offer effective structures to construct deep learning models, train and assess them. The suggested methodology is scalable, accurate and can be adapted to real world implementation.

10. RESULTS AND DISCUSSION

10.1 Comparison of the performances and accuracy analysis

The experimental findings prove that the proposed BiLSTM model is better than the traditional machine learning models, including Naive Bayes, Support Vector Machine, and Decision tree. As it is evident in Table 1, BiLSTM model is much more accurate than the base models are. In past research, deep learning models were reported to have accuracy above 90 percent whereas traditional models usually have an accuracy below 85 percent (Akter et al., 2025).

Table 1. Accuracy Comparison of Models

Model	Accuracy (%)
Naive Bayes	78
SVM	82
Decision Tree	80
BiLSTM	91

10.2 Precision, Recall, and F1-score Analysis

The evaluation metrics in Table 2 show that the BiLSTM model is more precise, recalls and F1-score than traditional methods. These measures illustrate the model capability of accurately detecting cases of cyberbullying with the least number of false positive outcomes and false negative outcomes..

Table 2. Performance Metrics Comparison

Model	Precision	Recall	F1-score
Naive Bayes	0.75	0.74	0.74
SVM	0.81	0.80	0.80
BiLSTM	0.90	0.89	0.89

10.3 Visualization and Interpretation of Results

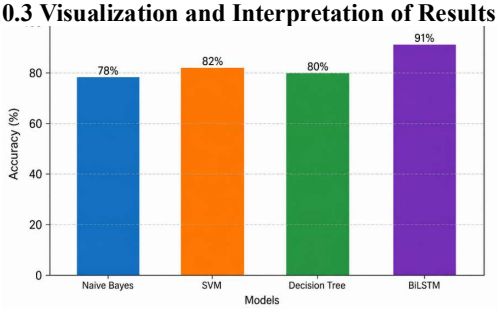

Figure 1. Model Accuracy Comparison

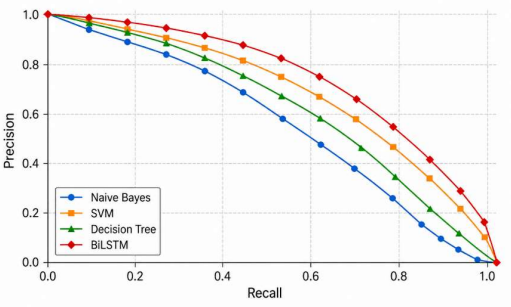

Figure 2. Precision-Recall Curve

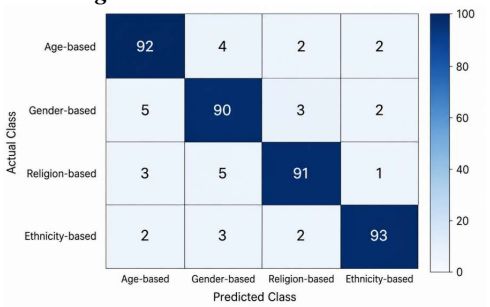

Figure 3. Confusion Matrix

The findings depicted in Figure 1, Figure 2 and Figure 3 indicate that the BiLSTM model has an overall higher performance in all the evaluation metrics. The confusion matrix indicates better categorization of cyberbullying types, whereas the curve of precision and recall indicates equal performance.

10.4 Case Analysis and Model Evaluation

As illustrated in Table 3, the model accurately points out instances of explicit and implicit cyberbullying. Nevertheless, there are certain misclassifications in very ambiguous sentences..

Table 3. Sample Classification Results

Text	Actual	Predicted
"You are worthless"	Bullying	Bullying
"Nice joke idiot"	Bullying	Non-Bullying

Additionally, **Table 4** presents a comparative analysis of model strengths and weaknesses.

Table 4. Model Strengths and Weaknesses

Aspect	Traditional Models	BiLSTM
Context Understanding	Low	High
Accuracy	Moderate	High
Scalability	Limited	High

Overall, the proposed model demonstrates improved performance, though challenges remain in handling ambiguous language.

11. CONCLUSION

The study conducted in this paper shows that deep learning methods, especially Bidirectional Long Short-Term Memory (BiLSTM) model, are effective in identifying cyberbullying in social media platforms. The research points to the weaknesses of classical machine learning methods, which do not reflect the contextual and semantic dependencies among textual data. The proposed system, incorporating Natural Language Processing techniques and word embeddings, is capable of improving the classification performance and accuracy.

The experimental findings confirm that the BiLSTM model is more accurate, precise, recall and F1-score compared to the traditional models. The fact that it can read both forward and backward, enables it to comprehend the intricate sentence structures and to detect subtle cases of cyberbullying effectively. This renders it a valid solution in practical application of it in real-life situations.

Moreover, the study underlines the significance of automated detection system in developing a safer online environment. As the amount of data on the social media increases, intelligent systems are needed to monitor and avert harmful interactions. The proposed solution leads to the improvement of the detection of cyberbullying as it offers a scalable and efficient solution. Comprehensively, the research proves that deep learning is a critical part in solving current challenges in internet communication and content control.

12. FUTURE WORK

The future in detecting cyberbullies can be based on the implementation of the latest deep learning models including Transformers and Bidirectional Encoder Representations of Transformers (BERT). These models have demonstrated outstanding performance in terms of understanding contextual and semantic relationships in text, which puts them at a good place to enhance detection accuracy. Recent developments have shown that transformer-based models are more effective than traditional deep learning-based models (Philipo et al., 2024).

The other direction is the elaboration of multilingual cyberbullying detection systems. Most of the models available are also limited to a particular language, and this limits their usability in different global environments. The multilingual datasets and language models can be used to increase the capacity of the system to detect cyberbullying in different cultures and languages.

It is also a priority area of work in the future because real-time deployment of detection systems is a priority area. The proposed model can be used with APIs of social networks and allow monitoring the content created by users continuously and taking

immediate action in case of abuse. Besides, the introduction of the continuous learning mechanisms will enable the model to respond to the changing patterns of the language and emergent forms of cyberbullying.

Overall, future advancements should focus on improving accuracy, scalability, and adaptability to ensure effective and reliable cyberbullying detection in dynamic online environments.

REFERENCES

1. Abdullah, A., Latif, I., Hafeez, N., Ullah, F., Sidorov, G., & Gelbukh, A. (2025). Cyberbullying detection on social media using machine learning techniques. https://www.researchgate.net/publication/396008271
2. Akter, M. S., et al. (2025). Cyberbullying detection on social media platforms. https://ijcaonline.org/archives/volume186/number61/akter-2025-ijca-924395.pdf
3. Akter, M. S., Shahriar, H., & Cuzzocrea, A. (2023). A trustable LSTM-autoencoder network for cyberbullying detection on social media using synthetic data. arXiv. https://arxiv.org/abs/2308.09722
4. Anuharini, N., Devadharshini, C. R., & Sowmiya, S. (2025). Social media cyberbullying detection applying machine learning. IEEE Conference Proceedings. https://www.researchgate.net/publication/392988535
5. Cuzzocrea, A., et al. (2025). Cyberbullying detection, prevention, and analysis on social media. https://www.mdpi.com/1999-5903/17/2/84
6. Fati, S. M., et al. (2023). Cyberbullying detection on Twitter using deep learning models. Mathematics, 11(16), 3567. https://www.mdpi.com/2227-7390/11/16/3567
7. Fitro, A., Wibowo, M. A., & Widodo, C. E. (2025). Automatic detection of cyberbullying on text, image, and video: A systematic literature review. https://www.researchgate.net/publication/398982073
8. Geetha, R. (2025). An efficient cyberbullying detection framework using hybrid deep learning models. https://www.inderscience.com/info/inarticle.php?artid=146156
9. Gowthami, S., & Rajesh, S. (2024). Cyber bullying detection using deep learning and natural language processing. Proceedings of ICICNIS 2024. https://www.researchgate.net/publication/387912659
10. Habiba, U., et al. (2025). Deep learning-based cyberbullying detection using BiLSTM. https://ijcaonline.org/archives/volume187/number21/habiba-2025-ijca-925105.pdf
11. Hasan, M. T., et al. (2023). A review on deep-learning-based cyberbullying detection. https://www.researchgate.net/publication/370682285
12. Hithnawi, R. I. (2025). Cyberbullying detection using BERT-based models. Cybersecurity Journal. https://academic.oup.com/cybersecurity/article/11/1/tyaf030/8313772
13. Hoque, M. N., et al. (2025). Advancing cyberbullying detection in low-resource languages. Frontiers in Artificial Intelligence. https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1679962/full
14. IJEDR. (2026). Machine learning-based cyberbullying detection system. https://rjwave.org/ijedr/papers/IJEDR2601399.pdf
15. IJRASET. (2025). Detection of cyberbullying using BiLSTM. https://www.ijraset.com/research-paper/detection-of-cyberbullying-using-bilstm
16. IJRPR. (2025). Cyberbullying detection on social media using machine learning. https://ijrpr.com/uploads/V6ISSUE3/IJRPR40317.pdf
17. IRJMS. (2024). Enhancing text classification with cyberbullying detection. https://www.irjms.com/wp-content/uploads/2024/10/Manuscript_IRJMS_01131_WS.pdf
18. Jaradat, G. (2025). Deep learning approaches for detecting cyberbullying on social media. Journal of Computational Communication Engineering. https://ojs.bonviewpress.com/index.php/JCCE/article/view/4162
19. Kumar, S., et al. (2025). Detecting cyberbullying in social media using NLP-based techniques. https://indjst.org/download-article.php?Article_Unique_Id=INDJST13943
20. Mazhar, A. A., et al. (2026). AI-powered detection of cyberbullying in short-form video platforms. PMC. https://pmc.ncbi.nlm.nih.gov/articles/PMC12893605/
21. Mohiuddin, G. M., et al. (2025). Deep learning models for culturally aware cyberbullying detection. Springer. https://link.springer.com/article/10.1007/s44163-025-00577-2

22. Philipo, A. G., et al. (2024). Assessing text classification methods for cyberbullying detection. https://arxiv.org/abs/2412.19928
23. Purkayastha, B. S., et al. (2025). Advancing cyberbullying detection using hybrid ML and DL frameworks. https://www.scitepress.org/Papers/2025/134362/134362.pdf
24. Ramadan, F., Hassan, E., & Omara, F. (2025). Cyberbullying detection using NLP techniques. https://www.researchgate.net/publication/401572963
25. Sakib, S. S., et al. (2024). Cyberbullying detection using transformer-based approaches for social media. Elsevier (Open Access Summary). https://doi.org/10.1016/j.nlp.2024.100104
26. Sree, S., & Joseph, N. (2026). Detection of cyberbullying in social media streams using temporal fusion networks. International Research Journal of Multidisciplinary Technovation. https://doi.org/10.54392/irjmt26216
27. Susmitha, V., Nagarani, J., & Lavanya, P. (2024). Detection of cyberbullying using machine learning algorithms. https://ijnrd.org/papers/IJNRD2405159.pdf
28. Unnava, S. (2024). A study of cyberbullying detection and classification techniques. https://etasr.com/index.php/ETASR/article/view/7621