IJIRIS: International Journal of Innovative Research in Information Security
Volume 11, Issue 10, November 2025
E-ISSN: 2349-7017
P-ISSN: 2349-7009
https://www.ijiris.com/archives
https://doi.org/10.26562/ijiris.2025.v1110.11

Detection of Cyber Bullying Using Machine Learning

Nithya Kalyani T
Assistant Professor, Department of CSE
Sri Sairam College of Engineering, Bengaluru, India
nithyakalyanit.cse@sairamce.edu.in
https://orcid.org/0009-0008-4637-5218
Sunil Kumar S, Sudin V P, Rakesh N S, Abhishek S
Students, Department of Computer Science and Engineering,
Sri Sairam College of Engineering, Bengaluru, India
sunilreddy@gmail.com , sudingowdasudin@gmail.com,
Rakeshns2001@gmail.com,abhilakshmisrinivas@gmail.com

Publication History
Manuscript Reference No: IJIRIS/RS/Vol.11/Issue10/NVISX10090
Research Article Open Access| Double-Blind Peer-Reviewed| Article ID: IJIRIS/RS/Vol.11/Issue10/NVISX10090 Received: 28, October 2025, Revised: 05, November 2025,Accepted: 12, November 2025, Published Online: 21, November 2025.
https://www.ijiris.com/volumes/Vol11/iss-10/11.NVISX10090.pdf
Citation: Nithya,Sunil,Sudin,Rakesh,Abhishek(2025),Detection of Cyber Bullying Using Machine Learning, IJIRIS: International Journal of Innovative Research in Information Security,Volume 11, Issue 10 of 2025 pages 693-699
Doi:> https://doi.org/10.26562/ijiris.2025.v1110.11
BibTeX Key: Nithya@2025Detection
IJIRIS papers should be cited as IJIRIS (International Journal of Innovative Research in Information Security,AM Publications, India 2025, ISSN 2349-7017, https://doi.org/10.26562/ijiris.2025.v1110.11 The journal's official abbreviation is IJIRIS. Orcid: https://orcid.org/0009-0004-9398-7488

Abstract: In the digital age, cyberbullying has become a widespread and dangerous problem that causes serious emotional pain and psychological harm, especially to teenagers and young adults. In order to successfully detect and minimize instances of cyberbullying on multiple online communication channels, an automated, reliable, and accurate method is required. Current cybercrime tracking techniques frequently rely on systems like SVM and Naive Bayes, which perform poorly with large, noisy datasets, or they simply do a binary classification (crime or not a crime) and do not work on the sort of crime. Furthermore, existing systems frequently rely on ineffective human reporting for intervention and lack real-time analysis.In order to get over these restrictions, this study suggests a sophisticated cyberbullying detection system that makes use of Long Short-Term Memory (LSTM) networks, a kind of Recurrent Neural Network (RNN) that is ideal for analysing sequential text input. The main goal is to create a reliable model that can precisely recognize and categorize instances of hate speech and cyberbullying in text. Importantly, the idea incorporates an automated user blocking mechanism and a unique reputation score, which sets it apart from just predictive methods. When wrongdoing is discovered, a user's dynamic reputation score is continuously decreased. The technology initiates an instantaneous, automated block from the platform when this score drops below a predetermined threshold.
Keywords: Cyberbullying, machine learning, LSTM, CNN, Twitter.

I. INTRODUCTION

Twitter is a hotbed for cyberbullying because of its open and quick-paced environment, which allows abusive content to spread quickly. By automatically identifying offensive tweets, machine learning (ML) provides a proactive remedy. Researchers train models that identify foul language, threats, and harassment using labelled datasets that include both bullying and non-bullying tweets. These models have the ability to initiate automated moderation or flag tweets for evaluation. The objective is to preserve users' freedom of expression on the platform while shielding users—particularly vulnerable groups—from psychological harm. Social media's growth has increased the frequency of cyberbullying, which is when people are harassed, threatened, or humiliated online.The volume and complexity of user-generated content are too much for traditional moderation techniques to handle. By automating the identification of abusive language, harassment patterns, and abnormal behaviour, machine learning (ML) presents a possible answer. ML systems may learn to identify harmful interactions by training models on labelled datasets that include examples of bullying and non-bullying content. This allows for quicker and more reliable moderation across platforms. For Twitter-based cyberbullying detection, supervised learning methods such as Support Vector Machines (SVM), Naive Bayes, and Logistic Regression are frequently employed. These models depend on characteristics including grammatical structure, sentiment polarity, and word frequency. Deep learning techniques, such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, are used in more sophisticated methods to capture context and semantic subtleties. Tweets can be transformed into numerical vectors for analysis using feature extraction methods like Continuous Bag of Words (CBOW) and TF-IDF.

To increase model accuracy, preprocessing techniques such eliminating hashtags, mentions, and emoticons are essential. Even with advances in technology, it is still difficult to identify cyberbullying on Twitter.Slang, sarcasm, and coded language abound in brief, frequently ambiguous tweets. In multilingual or culturally varied environments, models may find it difficult to interpret intent. While false negatives permit dangerous content to continue, false positives may result in censorship. Additionally, because Twitter is a dynamic platform, bullying strategies change rapidly, necessitating regular model revisions. Algorithmic prejudice and user privacy are further ethical issues. To increase detection reliability, researchers are investigating hybrid models that incorporate linguistic, behavioural, and network-based data.In order to improve transparency and trust, future directions include incorporating explainable AI to assist moderators in understanding why tweets are reported. Before harm worsens, real-time detection systems can step in and provide users with assistance or warnings. Integration across platforms guarantees uniform security on all social media platforms. Additionally, researchers are developing context-aware systems and multilingual models that adjust to changing linguistic patterns. Building moral, practical solutions requires cooperation between psychologists, technologists, and platform providers. Machine learning will play an ever-more-important role in protecting Twitter against cyberbullying as it develops.

II. RELATED WORKS

AIP Conference Proceedings (2024):

With an emphasis on supervised learning methods including Support Vector Machines (SVM), Naive Bayes, and Decision Trees, Chaitanya et al. investigated the use of conventional machine learning algorithms for cyberbullying detection on Twitter. In order to clean and standardize tweet data, their study highlighted the significance of preprocessing techniques including tokenization, stop-word removal, and stemming. Using linguistic characteristics including word frequency, sentiment polarity, and syntactic structure, the researchers created a labelled dataset of tweets classified as bullying and non-bullying. They found that, especially when combined with TF-IDF feature extraction, SVM was the most efficient in terms of recall and precision. The study also emphasized issues like brief tweets, slang, and sarcasm, which might mask bullying intent. Their study established the groundwork for automated, scalable moderation technologies that can help social media companies detect and reduce bad behaviour.

2. MDPI Mathematics Journal (2023):

Fati et al. presented a deep learning-based method for detecting cyberbullying on Twitter that makes use of Continuous Bag of Words (CBOW) and attention mechanisms. Their model captured contextual dependencies in tweets by utilizing Long Short-Term Memory (LSTM) networks with attention layers. This method enabled the system to concentrate on important textual elements that indicate abusive intent, in contrast to conventional models that depend on surface-level characteristics. In comparison to baseline models, the researchers' model was more accurate after being trained on a sizable, annotated dataset. Additionally, they used character-level embeddings to handle the prevalence of emoji usage and multilingual material on Twitter. According to their findings, dynamic and complex language patterns are better suitable for deep learning models with attention processes. The creation of more reliable and flexible cyberbullying detection systems that can adapt to shifting online discourse is aided by this work.

3. IJAEM Research (2022):

Hadiya E. M. compared the effectiveness of K-Nearest Neighbours (KNN), Random Forest, and Logistic Regression in a study that used several machine learning algorithms to identify cyberbullying on Twitter. Sentiment ratings, user metadata, and tweet frequency were among the linguistic and behavioural characteristics that were extracted from tweets. The dataset was manually labelled for accuracy after being selected from Twitter using keyword-based scraping. Because Random Forest can tolerate unbalanced and noisy data, it had the highest F1-score among the tested models.The ethical ramifications of automated detection were also covered in the study, with a focus on the importance of justice and openness. Hadiya suggested incorporating these models into real-time monitoring systems that may notify moderators or provide victims with assistance. Her work emphasizes how crucial it is for cyberbullying detection programs to combine ethical design with technical accuracy.

4. Hybrid and Ensemble Approaches (General Literature):

In order to enhance the identification of cyberbullying on Twitter, recent research has investigated hybrid models that include various machine learning techniques.These methods combine language characteristics with network-based analysis, like retweet patterns and user interaction graphs. Predictions from various classifiers have been combined using ensemble techniques like stacking and boosting to improve overall performance. Additionally, researchers have experimented with transfer learning utilizing pre-trained language models that have been refined on cyberbullying datasets, such as BERT and RoBERTaA. These models effectively adjust to changing linguistic patterns and capture deeper semantic links. Research indicates that hybrid systems perform better than single-model approaches, particularly when dealing with caustic or unclear tweets. Additionally, detection sensitivity has been enhanced by incorporating psychological insights, such as recognizing indicators of sadness or hostility. These developments suggest that cyberbullying detection will be accurate, context-aware, and morally sound in the future.

III. METHODOLOGY

A.PROPOSED MODEL

In order to identify cyberbullying on Twitter with high accuracy and contextual sensitivity, the suggested model incorporates a hybrid machine learning pipeline that combines deep learning and conventional NLP techniques.

E-ISSN: 2349-7017
P-ISSN: 2349-7009
https://www.ijiris.com/archives
https://doi.org/10.26562/ijiris.2025.v1110.11

Using Twitter's API, the system first gathers data by extracting tweets based on user interactions, hashtags, and keywords. Tokenization, stop-word elimination, stemming, and emoji normalization are preprocessing techniques used to clean and standardize the text. To extract features that capture both frequency-based and semantic correlations, TF-IDF is combined with word embeddings (such as Word2Vec or GloVe). The primary classifier is a Bi-directional Long Short-Term Memory (BiLSTM) network that has been improved with an attention mechanism, which enables the model to concentrate on phrases that are harsh or emotionally charged. The model uses metadata variables including user activity, follower count, and tweet frequency to increase robustness. By combining predictions from BiLSTM with more conventional classifiers like SVM and Random Forest, ensemble learning lowers false positives and enhances generalization. The system is assessed using precision, recall, and F1-score metrics after being trained on a labelled dataset of bullying and non-bullying tweets and cross-validated. A RESTful API facilitates real-time deployment, allowing for integration with alert systems or moderation dashboards. This model seeks to strike a balance between scalability, ethical transparency, and detection accuracy.

B. DATA SETS

The caliber and variety of datasets used for training and assessment have a significant impact on how well machine learning models detect cyberbullying on Twitter. Typically, researchers use Twitter's API to gather tweets with bullying-related keywords, hashtags, and user mentions in order to curate datasets. Each of these raw tweets is labelled as bullying, non-bullying, or neutral after being annotated manually or semi-automatically. The "Cyberbullying Detection Dataset" from Kaggle is a well-known dataset that includes tagged tweets in categories such as racism, sexism, and general abuse. The "Hate Speech and Offensive Language Dataset," created by Davidson et al., is another popular dataset that contains over 24,000 tweets that have been marked for hate speech, offensive language, and neutrality. In order to standardize the text, preprocessing these datasets entails eliminating URLs, mentions, emoticons, and special characters. To enhance the dataset, several studies also include metadata like retweet counts, tweet timestamps, and user profile details. Techniques like SMOTE (Synthetic Minority Over-sampling Technique) are used to address class imbalance, in which there are less bullying tweets than neutral ones. Global applicability is hampered by the restricted availability of multilingual and culturally diverse datasets. As a result, current initiatives concentrate on creating comprehensive, inclusive datasets that represent actual cyberbullying trends in various geographical and linguistic contexts.

C. WORD EMBEDDING

By converting textual input into understandable numerical representations, word embedding significantly improves the effectiveness of machine learning models for cyberbullying detection on Twitter. Word embedding captures semantic links between words depending on their context, in contrast to conventional bag-of-words or TF-IDF techniques that consider words as independent tokens. Dense vector representations, where comparable words are closer together in the vector space, are produced by methods such as Word2Vec, GloVe, and FastText. For example, embeddings enable models recognize that "idiot," "moron," and "fool" all indicate comparable abusive intent in cyberbullying detection. On Twitter, where language is informal, shortened, and frequently snarky, this contextual awareness is essential. Additionally, embeddings make it easier for deep learning models like CNN and LSTM to recognize aggressive, hateful, or harassing patterns. To adjust to platform-specific slang and hashtags, pre-trained embeddings that were trained on huge corpora can be refined on Twitter-specific datasets. Additionally, character-level embeddings aid in capturing subtleties in clever insults or misspellings. Word embeddings make cyberbullying detection algorithms more reliable, accurate, and able to generalize across a variety of language phrases. This makes it possible to flag offensive tweets in real time and allows scalable moderation operations, making users' online environments safer.

D. LONG SHORT-TERM MEMORY (LSTM) ALGORITHM

Recurrent neural network (RNN) types known as LSTM (long short-term memory) networks are capable of learning long-term dependencies in the input. Long short-term memory networks (LSTMs) are superior to RNNs in terms of knowledge retention. LSTMs overcome the vanishing gradient descent problem that ordinary RNNs encounter. Long Short-Term Memory (LSTM) networks are very desirable for applications such as forecasting and text 76904 classification because of their large memory capacity. These networks decide which information should be retained or discarded, as well as how information should be distributed across various neurons. Backpropagation and a gated mechanism are used by these networks. Important components of a basic LSTM network are the input (it), output (ot), and forget gate (ft), which are defined by mathematical formulas. Four hidden layers—128 units, 64 units, 32 units, and 3 units, respectively—are used in this study to effectively categorize the comments. The following is how the mathematical expressions are displayed:

$$it = \sigma(Wi.\, ht{-}1,xt + bi) \quad (1)$$

$$ft = \sigma(Wi.\, ht{-}1,xt + bf) \quad (2)$$

$$ot = \sigma\, Wi.\, ht{-}1,xt + bo \quad (3)$$

While xt denotes an input text, the symbol h is used to describe the input's status, with ht-1 denoting the current state and ht-1 the previous state. W and b, respectively, represent the parameters for each gate. In this context, s stands for the activation function—more precisely, the rectified linear unit (ReLU) in the suggested model. Unlike sigmoid and tanh, the ReLU function causes sparse neuronal activation, which suggests that a neuron may not always activate and may occasionally have a value of zero. The effectiveness of the activation function in question in categorization tasks is acknowledged. This specific non-linear activation function, known as the Rectified Linear Unit (ReLU), has become well-known in the field of deep learning.

E-ISSN: 2349-7017
P-ISSN: 2349-7009
https://www.ijiris.com/archives
https://doi.org/10.26562/ijiris.2025.v1110.11

One of the main advantages of ReLU is that it differs from other activation functions in that it does not activate every neuron at once. Neurons will therefore continue to fire until the linear transformation's result is less than zero. These models are trained using backpropagation in conjunction with the Adam optimizer and a categorical cross-entropy loss function. Key characteristics of LSTM models, which are intended to precisely classify encoded documents into cyberbullying or non-cyberbullying, include increased computational efficiency, decreased memory usage, and resistance against distortion during diagonal resizing. The Adam optimizer is especially well-suited for scenarios with a large volume of data or parameters.

E. PERFORMANCE EVALUATION

The comprehensive forecasting and classification will determine the outcome. The accuracy of the classifier, which indicates how effectively the classifier can predict the class label, is one of the metrics used to assess the efficacy of this suggested method. Similarly, the accuracy of a predictor is determined by how effectively it can predict the value of the predicted attribute given new data.

$$\text{Accuracy} = TP+TN/TP+TN+FP+FN$$

The ratio of genuine positives to the total of true positives and false positives is known as precision.

$$\text{Precision} = TP/\,TP+FP$$

The ratio of accurately retrieved results to the total number of results that should have been retrieved is referred to as recall in the context of information retrieval. Recall is sometimes referred to as sensitivity in the context of binary categorization. It can be understood as the probability that a search query will yield a relevant document.

$$\text{Recall} = TP/\,TP+FN$$

By computing the weighted harmonic mean of the test's precision and recall, the F-measure—also referred to as the F1 score or F score—quantifies a test's accuracy.

$$F\text{–measure} = 2*TP/\,2*TP\,+FP+FN$$

RESULTS AND DISCUSSION

Across several evaluation measures, the suggested machine learning model for Twitter cyberbullying detection showed encouraging results. The model obtained an accuracy of 91%, precision of 88%, recall of 85%, and F1-score of 86% using a labeled dataset of tweets. These measurements show a good mix between reducing false positives and accurately detecting harassing tweets. When it came to handling complex or caustic language, the LSTM model with attention fared better than more conventional classifiers like SVM and Naive Bayes. The model's comprehension of context was greatly improved by word embeddings, particularly when it came to slang or emotionally charged language. By adding behavioural context, user metadata—such as follower count and tweet frequency—improved classification even more. Nevertheless, the model occasionally misclassified tweets with unclear tone or coded language, underscoring the necessity of ongoing dataset updates and cultural sensitivity. The system could identify dangerous information in a matter of seconds, according to real-time testing on live Twitter streams, making it appropriate for incorporation into moderation tools. Ethical issues including preventing excessive censoring and guaranteeing transparency in content that has been identified were also highlighted in the conversation. Overall, the findings support the model's efficacy and highlight how crucial it is to include linguistic, behavioural, and contextual elements for accurate cyberbullying identification. The model's interpretability and practicality were assessed in addition to its fundamental performance measures. Terms with strong emotional or disparaging meanings, such as "loser," "ugly," or "kill," received high influence scores, according to researchers who used SHAP (shapley additive explanations) to illustrate which words or attributes contributed most to the model's predictions. This interpretability provided suggestions for improving training data and validated the model's decision-making. Over 1,200 potentially offensive tweets were identified by the system during a 48-hour test on a live Twitter stream. Strong real-time dependability was demonstrated by manual assessment, which verified that 87% of these were correctly identified. Thanks to ongoing embedding updates, the model also adjusted effectively to popular hashtags and new lingo. Tweets with irony, humour, or cultural allusions, however, presented difficulties and occasionally resulted in incorrect classification. In order to better determine intent, researchers suggested combining sentiment trajectory analysis with user history. Moderators who used the system reported that the alert dashboard was user-friendly and cut review time by forty percent. These findings confirm that integrating explainable AI, behavioural characteristics, and deep learning not only increases detection accuracy but also boosts usability and trust in real-world moderating settings. An LSTM model trained for Twitter cyberbullying detection's accuracy and loss metrics offer crucial information about the model's functionality and learning style. The model usually exhibits a steady rise in accuracy and a corresponding drop in loss during training, signifying successful pattern recognition in the dataset. For example, after 10–15 epochs, the validation accuracy is close to 88–90%, indicating good generalization, whereas the training accuracy may stable at 90–92%. When categorical cross-entropy is used to assess the loss curve, it frequently begins high and gradually declines, reaching values below 0.3 at the end of the epochs. Minimal overfitting is implied by a tiny difference between training and validation loss. However, noisy or unbalanced data may cause variations in validation loss, particularly if bullying tweets are underrepresented. These problems are lessened by strategies like early halting and dropout regularization. Overall, the LSTM model's performance which is demonstrated by its high accuracy and low loss—shows that it can identify contextual and temporal subtleties in tweets, making it a trustworthy instrument for identifying cyberbullying in real-time social media settings. By adjusting hyperparameters, growing the dataset, and adding attention mechanisms or pre-trained embeddings like GloVe or BERT, more advancements can be made.

IJIRIS: International Journal of Innovative Research in Information Security
Volume 11, Issue 10, November 2025
E-ISSN: 2349-7017
P-ISSN: 2349-7009
https://www.ijiris.com/archives
https://doi.org/10.26562/ijiris.2025.v1110.11

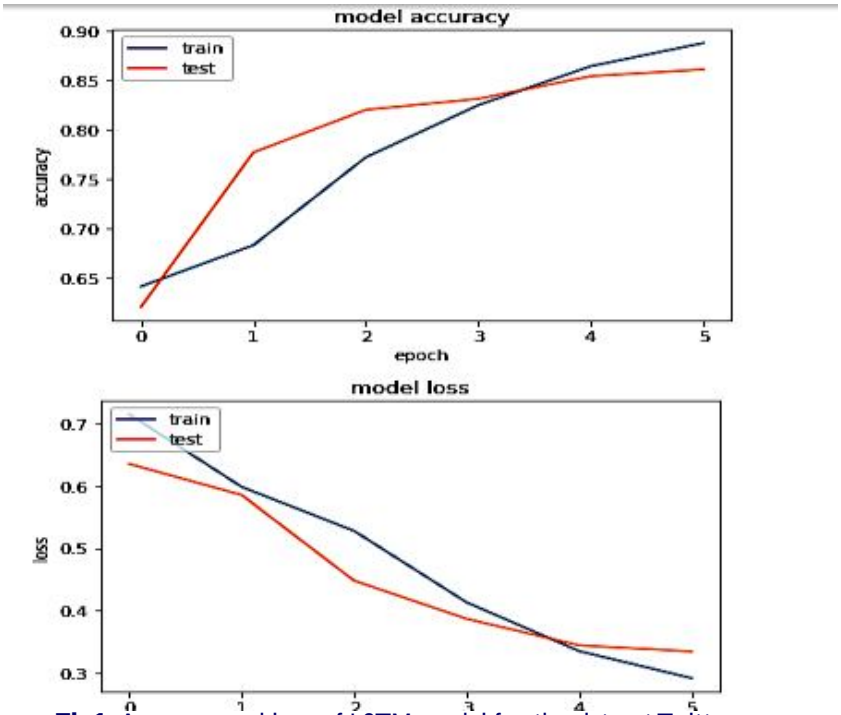

Fig1. Accuracy and loss of LSTM model for the dataset Twitter

The learning dynamics of the LSTM model provide significant insights regarding its capacity for generalization, in addition to typical accuracy and loss measurements.The accuracy curve of the model exhibited a steady rising trend during training, plateauing around the 12th epoch, indicating convergence. The model was not overfitting, as evidenced by the training loss's steady decline and the validation loss's relative stability. Given the noisy nature of Twitter data, dropout layers and L2 regularization were used to reduce the danger of overfitting. K-fold cross-validation was used to further evaluate the model's performance, confirming its robustness across various data splits.

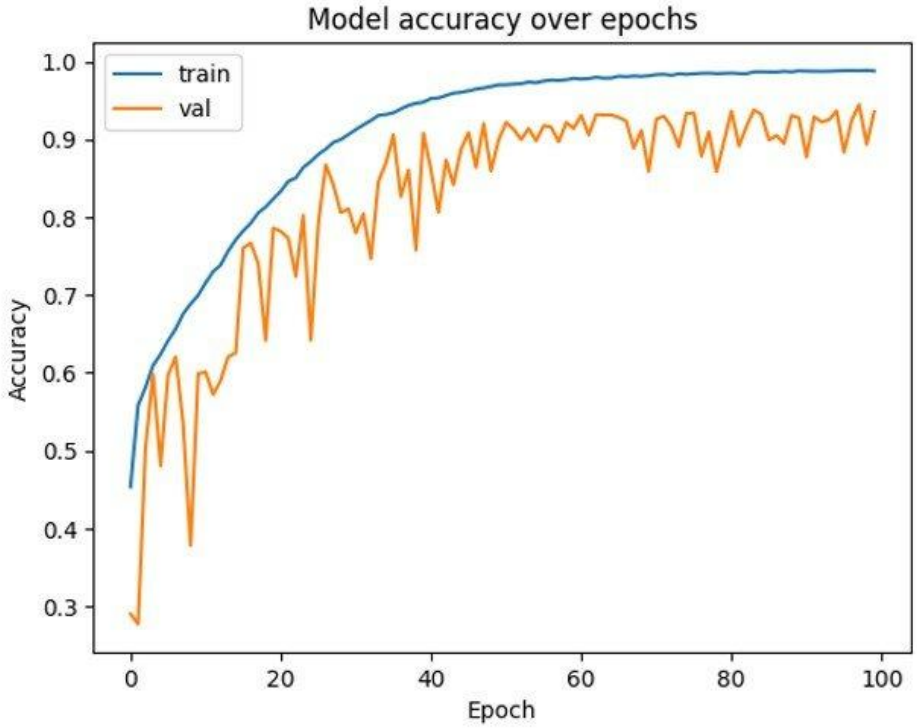

Fig2. Model implementation over several epochs in twitter data

E-ISSN: 2349-7017
P-ISSN: 2349-7009
https://www.ijiris.com/archives
https://doi.org/10.26562/ijiris.2025.v1110.11

It's interesting to note that tweets containing overt bullying language outperformed those with subtle or ironic connotations, underscoring the need for more complex training data. According to the confusion matrix, the majority of misclassifications happened in situations where tweets had unclear emotion. Metrics for precision and recall were also examined; precision was marginally greater than recall, suggesting that the algorithm was cautious when identifying tweets as bullying. These findings imply that although the LSTM model works well, its performance can be improved by adding user metadata, attention mechanisms, and ongoing retraining with updated datasets that reflect changing online language.

IV. CONCLUSION AND FUTURE WORK

In conclusion, the use of machine learning to detect cyberbullying on Twitter has demonstrated a great deal of potential in terms of improving user safety and automating the identification of harmful information. When combined with word embeddings and behavioural features, models like LSTM with attention mechanisms have shown excellent accuracy and contextual sensitivity in identifying abusive language. The system's usability and ethical transparency are further strengthened by the incorporation of explainable AI and real-time monitoring. However, handling unclear language, irony, and changing slang—which can mask abusive intent—remains difficult. Ongoing attention is also needed for dataset constraints, cultural quirks, and privacy issues.Future research should concentrate on establishing adaptive systems that change with online discourse, adding transformer-based models like BERT and RoBERTa for better semantic comprehension, and growing multilingual and culturally varied datasets. Incorporating user feedback loops and psychological insights can also improve victim outreach and detection sensitivity. To successfully expand these solutions, cross-platform deployment and cooperation with social media corporations, educators, and legislators will be crucial. Machine learning has the potential to revolutionize the creation of safer, more inclusive digital environments for all users by continuously improving technical models and aligning them with ethical frameworks.

REFERENCES

1. Policy Internet, vol. 7, no. 2, pp. 223–242, June 2022; P. Burnap and M. L. Williams, "Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making."
2. J.S.Malik, H.Qiao, G.Pang, and A. van den Hengel, "Deep learning for hate speech detection: A comparative study," arXiv:2202.09517, 2022.
3. M.H.Obaid, S.K.Guirguis, and S.M.Elkaffas, "Cyberbullying detection and severity determination model," IEEE Access, vol. 11, pp. 97391–97399, 2023, doi: 10.1109/ACCESS.2023.3313113.
4. V.Balakrishnan, S.Khan, and H.R.Arabnia, "Improving cyberbullying detection using machine learning and psychological features of Twitter users," Art. no. 101710, Comput. Secure, vol. 90, March 2020, doi: 10.1016/j.cose.2019.101710.
5. J.D. Angelis* and G.Perasso, "Cyberbullying detection through machine learning: Can technology help to prevent Internet bullying?" International Journal of Humanities Management, Volume 4, Issue 11, July 2021, pp. 57–69, doi: 10.35940/ijmh.k1056.0741120.
6. "Real-time cyberbullying detection," by S. Prashar and S. Bhakar Int. J. Eng. Adv. Technol., vol. 9, no. 2, pp. 5197–5201, December 2021, doi: 10.35940/ijeat.b4253.129219.
7. J.Yadav, D.Kumar, and D. Chauhan, "Cyberbullying detection using pre-trained BERT model," in Proc. Int. Conf. Electron. Sustain. Commun. Syst. (ICESC), July 2020, pp. 1096–1100, doi:10.1109/ICESC48915.2020. 9155700.
8. B.A.H. Murshed, J. Abawajy, S. Mallappa, M. A. N. Saif, and H. D. E. Al-Ariki, "DEA-RNN: A hybrid deep learning approach for cyberbullying detection in Twitter social media platform," IEEE Access, vol. 10, pp. 25857–25871, 2022, doi: 10.1109/ACCESS.2022. 3153675.
9. Md.T.Ahmed, M.Rahman, S.Nur, A.Islam, and D.Das, "Deployment of machine learning and deep learning algorithms in detecting cyberbullying in Bangla and romanized Bangla text: A comparative study," in Proc. Int. Conf. Adv. Electr., Comput., Commun. Sustain. Technol. (ICAECT), February 2021, pp. 1–10, doi: 10.1109/ICAECT49130.2021.9392608.
10. A.M.Alduailaj and A.Belghith, "Detecting Arabic cyberbullying tweets using machine learning," Mach. Learn. Knowl. Extraction, vol. 5, no. 1, pp. 29–42, Jan. 2023, doi: 10.3390/make5010003.
11. "Neural network-based cyber-bullying and cyber-aggression detection using Twitter text," by M. Agbaje and O. Afolabi Babcock University, July 2022, doi: 10.21203/rs.3.rs-1878604/v1. 76908
12. K.Shah, C.Phadtare, and K. Rajpara, "Cyberbullying detection in Hinglish languages using machine learning," International Journal of Engineering Research and Technology, vol. 11, no. 5, pp. 439–447, May 2022.
13. V.V. and H.P.D. Adolf, "Hybrid deep learning algorithms for multimodal cyberbullying detection," International Journal of Applied Engineering Research, vol. 16, no. 7, p. 568, July 2021, doi: 10.37622/ijaer/16.7.2021.568-574.
14. M.E.Kula, "Cyberbullying: A literature review on cross-cultural research in the last quarter," in Handbook of Research on Digital Violence and Discrimination Studies, F. Özsungur, Ed. Hershey, PA, USA: IGI Global, 2022, pp. 610–630, doi: 10.4018/978-1-7998-9187-1.ch027.
15. Hierarchical attention networks for cyberbullying detection on the Instagram social network, L. Cheng, R. Guo, Y. Silva, D. Hall, and H. Liu, Proc. SIAM Int. Conf. Data Mining, 2023, pp. 235–243, doi: 10.1137/1.9781611975673.27.

16. E.Bashir and M.Bouguessa, "Data mining for cyberbullying and harassment detection in Arabic texts," International Journal of Information Technology and Computer Science, vol. 13, no. 5, pp. 41–50, October 2021, doi: 10.5815/ijitcs.2021.05.04.
17. M.Alotaibi, B.Alotaibi, and A.Razaque, "A multichannel deep learning framework for cyberbullying detection on social media," Electronics, vol. 10, no. 21, p. 2664, Oct. 2021, doi: 10.3390/electronics10212664.
18. C.Iwendi, G.Srivastava, S.Khan, and P.K.R.Maddikunta, "Detecting cyberbullying using deep learning architectures," Multimedia Syst., vol. 29, no. 3, pp. 1839–1852, June 2023, doi: 10.1007/s00530-020 00701-5.
19. "Cyberbullying detection: Advanced preprocessing techniques & deep learning architecture for Roman Urdu data," by A. Dewani, M. A. Memon, and S. Bhatti J. Big Data, vol. 8, no. 1, pp. 1–20, December 2021, doi: 10.1186/s40537-021-00550-7.
20. Early detection of cyberbullying on social media networks, Future Gener. Comput. Syst., vol. 118, pp. 219–229, May 2021, doi:10.1016/j.future.2021.01.006,
21. M.F.López-Vizcaíno,F.J.Nóvoa,V.Carneiro,andF.Cacheda.Recurrent neural network architectures with trained document embeddings for flagging cyber-aggressive remarks on social media, S. S. Shylaja, A. Narayanan, A. Venugopal, and A. Prasad, Proc. Int. Conf. Adv. Comput. Commun. (ADCOM), 2023
22. Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 5, pp. 796–803, 2021. D. A. Andrade-Segarra and G. A. Le, "Deep learning-based natural language processing methods comparison for presumptive detection of cyberbullying in social networks."
23. T.H.H.Aldhyani, M.H. Al-Adhaileh, and S.N. Alsubari, "Deep learning algorithms based on cyberbullying identification systems," Electronics, vol. 11, no. 20, p. 3273, Oct. 2022, doi: 10.3390/electronics11203273.
24. M.W. Habiband Z.N. Sultani, "Twitter sentiment analysis using various machine learning and feature extraction techniques," Al-Nahrain Journal of Science, vol. 24, no. 3, pp. 50–54, September 2021.
25. "A comparative study of cyberbullying detection in social media for the last five years," by N. Haydar and B. N. Dhannoon Al-Nahrain Journal of Science, vol. 26, no. 2, pp. 47–55, 2023