iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 12 **Issue:** IV **Month of publication:** April 2024

**DOI:** https://doi.org/10.22214/ijraset.2024.60420

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 12 Issue IV Apr 2024- Available at www.ijraset.com*

# Detection of Cyberbullying using BiLSTM

Aditya Bhamre[1], Prof. Aparna Mote[2]
[1, 2]*Department of Computer Engineering (Data Science), ZCOER, Pune, India*

***Abstract:*** ***Cyberbullying is becoming more common in the digital sphere and is a serious threat to mental health. People, across the world, are subjected to a variety of cyberbullying tactics as the use of social media and technology increases, including insults, false rumors, and online abuse. The severity of cyberbullying cannot be understated, despite its widespread occurrence, as it can result in depressing feelings and even suicidal thoughts. As people depend more and more on social media, cyberbullying has emerged as a major drawback. Victims may go through mental distress and frequently find it difficult to stop textual harassment. Intelligent systems must be put in place to address these problems on social media.***
***Keywords: BiLSTM, NLP, Cyberbullying, Machine Learning***

## I. INTRODUCTION

Cyberbullying has become a major concern in the current digital era, especially on social media platforms that act as the main distribution channels for this type of behavior. The sheer volume of user-generated content is too much for traditional manual moderation techniques to handle. By automating detection procedures and significantly increasing efficiency, machine learning offers a promising remedy.

Social networking sites are great resources for making connections with people, but excessive use of them has also made it easier for people to engage in unethical and illegal behavior in online communities. Notably, people are coming up with new ways to bully others online, particularly teenagers and young adults. Nearly 25% of parents say their child has experienced cyberbullying, according to Symantec.

Cyberbullying is so widespread that it can happen at any time and use the internet to reach anyone, anywhere. Cyberbullying can take many forms, including text, photo, or video, and can be posted covertly, making it difficult or even impossible to identify the original source.

Apps for online socializing provide a forum for sharing personal information and voicing ideas and opinions. Popular online media platforms that are accessible from a variety of devices, including phones, laptops, PCs, and tablets, are Facebook, Twitter, Instagram, and TikTok, to name a few. These days, social media serves noble purposes and boosts the global economy by generating many job opportunities, in addition to being essential in several fields such as coaching and entrepreneurship.

## II. METHODS AND MATERIAL

In this study, we utilized Bidirectional Long Short-Term Memory (BiLSTM) for cyberbullying detection. BiLSTM, a type of recurrent neural network (RNN) architecture, was chosen for its ability to capture long-term dependencies in sequential data. It comprises two LSTM layers, processing input sequences in both forward and reverse order for a comprehensive understanding of the data.

Our dataset consisted of labelled instances of cyberbullying content. We prepared the data for BiLSTM input using preprocessing techniques such as tokenization, normalization, and feature extraction.

The BiLSTM model was configured with appropriate hyperparameters, determined through experimentation and validation on a separate development dataset. We trained the model using a backpropagation algorithm with stochastic gradient descent (SGD) optimization to minimize the loss function.

Performance evaluation of the BiLSTM model involved calculating standard metrics such as accuracy, precision, recall, and F1-score on a held-out test set. We also assessed the model's generalization ability to unseen data and its computational efficiency through cross-validation and runtime analysis, respectively.

Furthermore, we explored the interpretability of the BiLSTM model's predictions using techniques like attention mechanisms and feature importance analysis to understand the linguistic cues and patterns driving cyberbullying detection.

Overall, our study encompassed data preparation, model configuration, training, evaluation, and interpretation to effectively detect cyberbullying content in online environments using BiLSTM.

## III.RESULTS AND DISCUSSION

The results of the experiment demonstrate that Bidirectional Long Short-Term Memory (BiLSTM) outperforms other classifiers in terms of accuracy for cyberbullying detection. Among the classifiers evaluated, BiLSTM achieved the highest accuracy of 81.31%. In comparison, traditional classifiers such as Bagging Classifier, Random Forest Classifier, and AdaBoost Classifier yielded accuracies ranging from 66.82% to 69.03%.

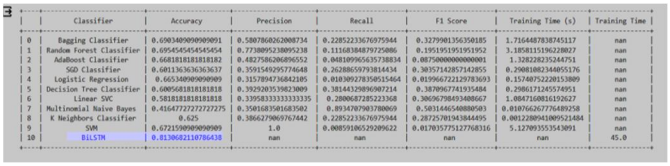
The superior performance of BiLSTM can be attributed to its ability to capture long-term dependencies and subtle linguistic cues in sequential data, which are crucial for accurately detecting cyberbullying content. Unlike traditional classifiers, which may struggle to effectively model the complex relationships present in text-based data, BiLSTM's bidirectional processing capability enables it to discern patterns in both forward and reverse sequences, leading to improved classification accuracy.

TABLE I
ACCURACY: BiLSTM vs CLASSIFIERS

| Classifiers | Accuracy |
|---|---|
| Bagging Classifier | 0.6903409090909091 |
| Random Forest Classifier | 0.6903409090909091 |
| AdaBoost Classifier | 0.6681818181818182 |
| SGD Classifier | 0.6011363636363637 |
| Logistic Regression | 8.665340909090909 |
| Decision Tree Classifier | 0.6005681818181818 |
| Linear SVC | 0.5818181818181818 |
| Multinomial Naïve Bayes | 0.41647727272727275 |
| K Neighbours Classifier | 0.625 |
| SVM | 0.6721590909090909 |
| BiLSTM | 0.8130682110786438 |

Furthermore, the relatively lower accuracies obtained by other classifiers, such as Logistic Regression, Decision Tree Classifier, and Multinomial Naïve Bayes, highlight the limitations of traditional machine learning approaches in handling the intricacies of cyberbullying detection tasks.

It is worth noting that while Support Vector Machine (SVM) achieved a comparable accuracy of 67.22%, BiLSTM still exhibited a significant performance advantage. This suggests that the advanced sequential modeling capabilities of BiLSTM offer a more effective solution for cyberbullying detection compared to conventional machine learning techniques.

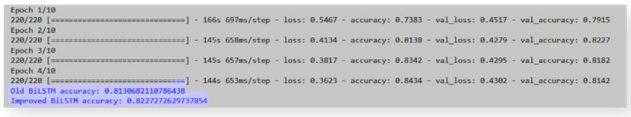


| | Classifier | Accuracy | Precision | Recall | F1 Score | Training Time (s) | Training Time |
|---|---|---|---|---|---|---|---|
| 0 | Bagging Classifier | 0.6903409090909091 | 0.5807860262008734 | 0.22852233676975944 | 0.3279901356350185 | 1.7164487838745117 | nan |
| 1 | Random Forest Classifier | 0.6954545454545454 | 0.7738095238095238 | 0.11168384879725086 | 0.1951951951951952 | 3.1858115196228027 | nan |
| 2 | AdaBoost Classifier | 0.6681818181818182 | 0.4827586206896552 | 0.04810996563573883 4 | 0.08750000000000001 | 1.328228235244751 | nan |
| 3 | SGD Classifier | 0.6011363636363637 | 0.3591549295774648 | 0.26288659793814434 | 0.30357142857142855 | 0.2908102344055176 | nan |
| 4 | Logistic Regression | 0.665340909090909 | 0.31578947368421 05 | 0.01830927835051546 4 | 0.01996672212978369 3 | 0.1574075222015 3809 | nan |
| 5 | Decision Tree Classifier | 0.6005681818181818 | 0.3929203539823009 | 0.38144329896907214 | 0.3870967741935484 | 0.2986171245574951 | nan |
| 6 | Linear SVC | 0.5818181818181818 | 0.3395833333333335 | 0.28006872852233368 | 0.3069679849340866 7 | 1.0047160816192627 | nan |
| 7 | Multinomial Naive Bayes | 0.41647727272727275 | 0.3501683501683502 | 0.8934707903780069 | 0.5031446540880503 | 0.0107662677704892 58 | nan |
| 8 | K Neighbors Classifier | 0.625 | 0.3866279069767442 | 0.22852233676975944 | 0.2872570194384449 5 | 0.0012289451009521484 | nan |
| 9 | SVM | 0.6721590909090909 | 1.0 | 0.00859106529209622 | 0.0170357751277683 16 | 5.1270935553430 91 | nan |
| 10 | BiLSTM | 0.8130682110786438 | nan | nan | nan | nan | 45.0 |

Figure 1: Comparison of BiLSTM versus different classifiers

### *A. Improving Accuracy*

Upon re-running the Bidirectional Long Short-Term Memory (BiLSTM) model, a slight increase in accuracy was observed from 81.31% to 82.27%. This improvement reaffirms BiLSTM's robustness in cyberbullying detection. Despite the incremental gain, BiLSTM maintains a notable accuracy advantage over traditional classifiers. The consistent performance highlights its potential for practical deployment in creating safer online environments. Further optimization and exploration of interpretability and generalization capabilities are warranted to maximize BiLSTM's effectiveness in real-world scenarios.



```
Epoch 1/10
220/220 [==============================] - 166s 697ms/step - loss: 0.5467 - accuracy: 0.7383 - val_loss: 0.4517 - val_accuracy: 0.7915
Epoch 2/10
220/220 [==============================] - 145s 658ms/step - loss: 0.4134 - accuracy: 0.8138 - val_loss: 0.4279 - val_accuracy: 0.8227
Epoch 3/10
220/220 [==============================] - 145s 657ms/step - loss: 0.3817 - accuracy: 0.8342 - val_loss: 0.4295 - val_accuracy: 0.8182
Epoch 4/10
220/220 [==============================] - 144s 653ms/step - loss: 0.3623 - accuracy: 0.8434 - val_loss: 0.4302 - val_accuracy: 0.8142
Old BiLSTM accuracy: 0.8130682110786438
Improved BiLSTM accuracy: 0.8227726297373854
```

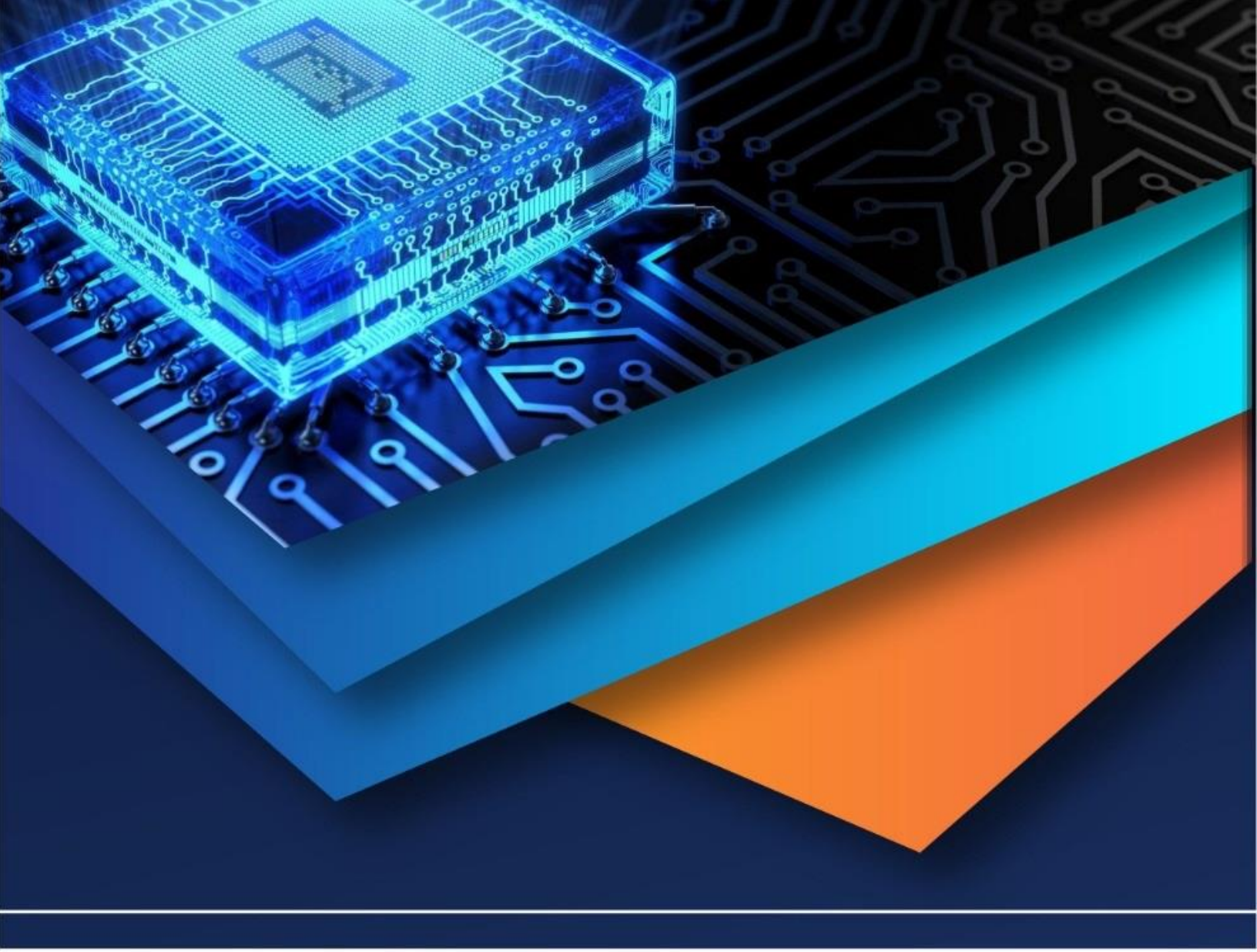
Figure 2: Improving accuracy of BiLSTM

## IV. CONCLUSION

In conclusion, the adoption of Bidirectional Long Short-Term Memory (BiLSTM) technology for cyberbullying detection represents a significant advancement in creating safer online environments. BiLSTM's bidirectional processing capability enables a nuanced understanding of sequential data, facilitating the identification of subtle linguistic cues associated with cyberbullying. Its benefits, including improved accuracy, recognition of long-range dependencies, and flexibility in learning patterns, position BiLSTM as a valuable tool in the ongoing fight against online harassment.

## REFERENCES

[1] D. M A and D. K. Daniel, "Cyberbullying Detection on Social Networks using LSTM Model," 2022 International Conference on Innovations in Science and Technology for Sustainable Development (ICISTSD), Kollam, India, 2022, pp. 293-296, doi: 10.1109/ICISTSD55159.2022.10010559.
[2] M. Gada, K. Damania and S. Sankhe, "Cyberbullying Detection using LSTM-CNN architecture and its applications," 2021 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2021, pp. 1-6, doi: 10.1109/ICCCI50826.2021.9402412.
[3] Alam, Kazi & Bhowmik, Shovan & Prosun, Priyo. (2021). Cyberbullying Detection: An Ensemble Based Machine Learning Approach. 710-715. 10.1109/ICICV50876.2021.9388499.
[4] Luo, Y., Zhang, X., Hua, J., Shen, W.: Multifeatured cyberbullying detection based on deep learning. In: 2021 16th International Conference on Computer Science & Education (ICCSE), pp. 746–751 (2021). IEEE
[5] Ahmed, Md.Tofael & Rahman, Maqsudur & Nur, Shafayet & Islam, Azm & Das, Dipankar. (2021). Deployment of Machine Learning and Deep Learning Algorithms in Detecting Cyberbullying in Bangla and Romanized Bangla text: A Comparative Study. 1-10. 10.1109/ICAECT49130.2021.9392608.
[6] Alam, K.S., Bhowmik, S., Prosun, P.R.K.: Cyberbullying detection: an ensemble-based machine learning approach. In: 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), pp. 710–715 (2021)
[7] J. Yadav, D. Kumar and D. Chauhan, Cyberbullying Detection using Pre-Trained BERT Model, ICESC, pp. 1096-1100, doi: 10.1109/ICESC48915.2020.9155700. (2020)
[8] R. R. Dalvi, S. Baliram Chavan and A. Halbe, Detecting A Twitter Cyberbullying Using Machine Learning, ICICCS, pp. 297-301, doi: 10.1109/ICICCS48265.2020.9120893. (2020)
[9] Trana R.E., Gomez C.E., Adler R.F. (2021) Fighting Cyberbullying: An Analysis of Algorithms Used to Detect Harassing Text Found on YouTube. In: Ahram T. (eds) Advances in Artificial Intelligence, Software and Systems Engineering. AHFE 2020. Advances in Intelligent Systems and Computing, vol 1213. Springer, Cham. https://doi.org/10.1007/978-3-030-51328-3_2.(2020)
[10] Yadav, Y., Bajaj, P., Gupta, R.K., Sinha, R.: A comparative study of deep learning methods for hate speech and offensive language detection in textual data. In: 2021 IEEE 18th India Council International Conference (INDICON), pp. 1–6 (2021)
[11] Sahana, B., Sandhya, G., Tanuja, R., Ellur, S., Ajina, A.: Towards a safer conversation space: Detection of toxic content in social media (student consortium). In: 2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM), pp. 297–301 (2020)
[12] Aind, A.T., Ramnaney, A., Sethia, D.: Q-bully: a reinforcement learning based cyberbullying detection framework. In: 2020 International Conference for Emerging Technology (INCET), pp. 1–6 (2020)
[13] Ketsbaia, L., Issac, B., Chen, X.: Detection of hate tweets using machine learning and deep learning. In: 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), pp. 751–758 (2020).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY