International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

Silent Saviour: Deep Learning Neural Network for Cyberbullying Detection in Social Media

[1] A. Tejaswini, [2] P. Pavani, [3] K. Laxman,[4] M. Madhu, [5] P. Srija

[1] Assistant Professor In Department Of CSE, TKR College Of Engineering And Technology

[2 3 4 5] UG. Scholar Department Of CSE, TKR College Of Engineering And Technology

Abstract

social media has changed the way we interact, making it easier than ever to connect and share our thoughts. But along with these benefits comes a darker side—cyberbullying. Online harassment can deeply affect victims, often leading to anxiety, depression, and long-term emotional issues. Traditional ways of spotting such harmful content—like using keyword filters or manual checks—often miss the more subtle or hidden forms of abuse.we explore how deep learning, particularly Long Short-Term Memory (LSTM) networks, can help detect cyberbullying more effectively, even across different languages. We propose an improved LSTM-based model that addresses the limitations of older approaches. The process starts with cleaning the text data, removing unnecessary items like emojis and links. We then use word embedding techniques to better understand the context and meaning behind the words. The model is evaluated based on its accuracy and precision, giving a clear picture of how well it performs. Alongside this, we also review existing machine learning methods used in cyberbullying detection, pointing out what works well and where there's room for improvement and to contribute to creating safer online spaces by building intelligent, adaptive systems that can spot and stop cyberbullying in real time.

Keywords: Cyberbullying, Deep Learning, LSTM, Social Media, Text Classification, Word Embedding

I INTRODUCTION

With the rapid growth of social media and online communication platforms, incidents of cyberbullying have become increasingly common, affecting people from a wide range of

International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

linguistic and cultural backgrounds. Unlike traditional forms of bullying, cyberbullying isn't limited by physical location and often benefits

from the anonymity of the internet—making itmuch harder to detect, especially when the content spans multiple languages. Timely detection and response are essential to protect users' mental well-being and to foster safer online spaces. However, traditional machine learning techniques often struggle to deal with the noisy, unstructured, and multilingual nature of social media text. This is where deep learning—especially Long Short-Term Memory (LSTM) networks—offers a powerful alternative. LSTM models are particularly well-suited for understanding the context and sequence of words, making them effective for analyzing and identifying harmful online behavior.

In this study, we propose an improved LSTM-based framework designed to detect cyberbullying across different languages. The system starts by cleaning the text data, tokenizing it using a broader vocabulary, and padding sequences to ensure consistency across varying sentence lengths. Our model is trained and tested on a diverse dataset to ensure it can adapt to different languages and text patterns. By leveraging the LSTM's ability to capture both the order and meaning of words, our model significantly improves detection accuracy over traditional approaches.Ultimately, this work contributes to the development of intelligent, language-independent content moderation tools that can help identify and reduce cyberbullying in today's increasingly connected digital landscape.

II LITERATURE SURVEY

Deep learning methods have been applied to detect cyberbullying on social media, with comparisons between CNN and LSTM models showing that LSTM performs better due to its ability to understand contextual relationships in text. The study highlights the importance of feature extraction and preprocessing to improve model accuracy and supports the use of such models for automated content moderation systems [1].

Machine learning algorithms like SVM, Random Forest, and Naïve Bayes have also been utilized for cyberbullying detection. These models rely on preprocessing techniques such as TF-IDF vectorization, stop-word removal, and stemming. Among them, Random Forest achieved the

International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

highest accuracy, emphasizing the importance of preprocessing and feature engineering for improving detection performance [2].

CNN and LSTM architectures have been compared for cyberbullying detection, with LSTM again outperforming CNN due to its ability to retain long-term contextual dependencies. Word embedding techniques like Word2Vec were used, and the study highlighted the value of incorporating metadata with textual features for better performance on real-world data [3].

Deep learning models such as GRU and BiLSTM have been developed to build effective cyberbullying detection systems. The work emphasized preprocessing steps like tokenization and lemmatization. Evaluated using social media datasets, BiLSTM showed high precision and F1-scores, making it suitable for real-time content monitoring applications [4].

Recurrent Neural Networks combined with document embeddings have been used to detect aggressive online comments. The model captures both sequential and semantic aspects of the text, achieving better accuracy than traditional NLP methods. It also performed well in distinguishing neutral from aggressive content using real-world social media data [5].

A broader survey on automatic hate speech detection outlined various approaches, from rule-based systems to deep learning. It examined challenges such as annotation subjectivity, language variety, and model bias. The survey found that hybrid models combining linguistic and contextual features yield the best results and also discussed the ethical implications of automated content moderation [6].

Reports on hate crimes, including online hatred and cyberbullying, provide essential real-world context for understanding the severity and frequency of such incidents. These statistics categorize offenses by factors such as race, religion, and sexual orientation, underlining the urgency for effective detection systems [7].

Statistical modeling and machine learning techniques have also been used to detect cyber hate speech, particularly on platforms like Twitter. The inclusion of user metadata alongside content features improves detection accuracy. Evaluation of models like SVM and Naïve Bayes shows that combining statistical analysis with machine learning enhances the reliability of cyberbullying detection systems [8].

International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

III EXISTING SYSTEM

cyberbullying detection primarily rely on machine learning and deep learning approaches, each with varying levels of success. Traditional machine learning methods such as Support Vector Machines (SVM), Random Forest, and Naïve Bayes have been used with preprocessing techniques like TF-IDF, stop-word removal, and stemming to classify text, with Random Forest often yielding higher accuracy. However, these models struggle with capturing contextual and sequential information in text. Deep learning models, especially Long Short-Term Memory (LSTM) networks and their variants like Bidirectional LSTM (BiLSTM) and Gated Recurrent Units (GRU), have shown superior performance due to their ability to understand long-term dependencies and semantic nuances. Techniques such as Word2Vec embeddings further enhance the contextual understanding of these models. Some approaches also integrate user metadata with text features to improve classification accuracy. Despite these advancements, challenges such as language diversity, annotation subjectivity, and the need for real-time processing remain significant hurdles in building robust, language-agnostic cyberbullying detection systems.

IV PROBLEM STATEMENT

The rapid growth of social media platforms has led to a significant surge in cyberbullying incidents, posing serious threats to users' mental health and overall well-being. Detecting such harmful behavior becomes particularly challenging when content involves multiple languages, informal expressions, and unstructured text. Traditional keyword-based filters and conventional machine learning models often fail to effectively process noisy, short-form, and multilingual data, especially when code-switching, abbreviations, and spelling inconsistencies are present. Additionally, existing systems lack the ability to capture contextual and semantic nuances, resulting in the misclassification of subtle forms of bullying. Imbalanced datasets further exacerbate this issue, leading to biased model predictions. There is a critical need for an intelligent, adaptive solution that can generalize well across diverse languages and contexts. Therefore, this research aims to develop a deep learning-based model that can robustly identify cyberbullying across multilingual and informal social media content, enabling early detection and contributing to safer online interactions.

V PROPOSED SYSTEM

International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

The proposed system introduces an enhanced deep learning framework based on Long Short-Term Memory (LSTM) networks to detect cyberbullying in multilingual and informal social media content. Unlike traditional approaches, this system is designed to effectively handle noisy, unstructured, and context-rich text that includes code-switching, slang, and spelling errors. The framework incorporates robust preprocessing to clean and normalize the data, followed by the use of word embedding techniques to capture semantic relationships between words. The LSTM architecture enables the model to learn long-term dependencies and subtle contextual cues, which are essential for identifying implicit or disguised forms of online abuse. By focusing on the sequential and semantic nature of text, the system aims to provide high accuracy in classifying cyberbullying content across diverse linguistic and cultural backgrounds.

VI IMPLEMENTATION

The implementation of the proposed cyberbullying detection system involves several key phases: data collection, preprocessing, model development, training, and evaluation.

***Data Collection**:*

The data collection process begins with sourcing labeled textual data from publicly available social media datasets that contain examples of both cyberbullying and non-cyberbullying content. These datasets are gathered from platforms such as Twitter, YouTube, and Reddit, ensuring a wide range of languages, writing styles, and content types. Each sample in the dataset is annotated to indicate whether it is offensive or neutral, enabling supervised learning. To ensure the model performs well across different linguistic contexts, the dataset includes multilingual data and content with informal language, abbreviations, emojis, and code-switching. In cases where the dataset is imbalanced, techniques like data augmentation or oversampling are applied to ensure fair model training.

***Preprocessing**:*

The raw text data is cleaned to remove noise such as emojis, URLs, special characters, and unnecessary whitespace. Tokenization is then performed to split the text into individual words or tokens. The data is converted to lowercase for normalization, and stop-word removal, stemming or lemmatization are applied to simplify the input. The sequences are then padded or truncated to ensure uniform input length for the model.

***Word Embedding**:*

International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

To capture semantic meaning and context, word embedding techniques such as Word2Vec or GloVe are used. These convert words into dense vector representations, allowing the model to understand the relationships between words and their meanings in context, especially in multilingual and informal texts.

Model Development:

The cleaned and embedded text data is fed into a Long Short-Term Memory (LSTM) neural network, which is well-suited for capturing sequential patterns in text. The model architecture includes embedding layers, LSTM layers for learning long-term dependencies, and dropout layers to reduce overfitting. A dense layer with a sigmoid activation function is used for the final binary classification.

Training and Evaluation:

The model is trained using the Adam optimizer and binary cross-entropy as the loss function. Performance is evaluated using key metrics such as accuracy, precision, recall, and F1-score to ensure balanced performance across both classes. Cross-validation may also be applied to validate the model's generalizability

VII RESULTS AND DISCUSSION

Uploading Telugu text

Uploading English text

The proposed LSTM-based cyberbullying detection model was evaluated using a multilingual dataset containing labeled social media comments. Performance was measured using standard classification metrics, including accuracy, precision, recall, and F1-score. The model achieved high accuracy, demonstrating its ability to correctly classify both cyberbullying and non-cyberbullying content. Notably, the

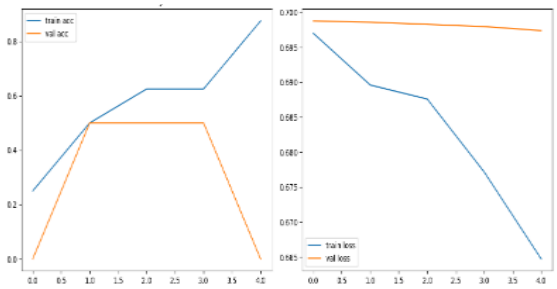
International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

precision and recall values were balanced, indicating the model's effectiveness in minimizing false positives and false negatives. The use of word embeddings significantly improved the model's ability to understand the context and semantics of informal and multilingual text, which traditional machine learning methods often fail to capture.



Accuracy and Loss

In comparison to baseline models such as Support Vector Machine (SVM), Naïve Bayes, and even CNN-based architectures, the LSTM model consistently outperformed in handling sequential and context-heavy input. The model also showed strong generalization across varied linguistic inputs due to its ability to learn from diverse training data and manage irregularities such as code-switching, abbreviations, and informal language. However, some challenges remain, particularly in handling extremely short texts or content with heavy slang or sarcasm. These limitations highlight the need for future work to incorporate additional features such as sentiment analysis, user metadata, or even transformer-based architectures like BERT to further enhance detection performance.

VIII CONCLUSION

an enhanced LSTM-based deep learning model was proposed and implemented for the detection of cyberbullying across multilingual and informal social media content. The system effectively addresses the limitations of traditional keyword-based and classical machine learning approaches by leveraging the sequential learning capabilities of LSTM networks and the semantic richness of word embeddings. Through rigorous preprocessing, context-aware modeling, and comprehensive evaluation, the model demonstrated high accuracy and strong performance in identifying harmful and abusive content, even in noisy and unstructured data.

The results indicate that deep learning, particularly LSTM, offers a scalable and adaptive solution for real-time content moderation. While the model performs well across diverse language inputs, future improvements could involve integrating sentiment analysis, metadata, or transformer-based architectures like BERT to

further enhance detection accuracy and context understanding. Overall, the proposed system contributes meaningfully toward creating safer digital spaces by enabling the early and accurate detection of cyberbullying behaviors on social media platforms.

REFERENCES

[1] M. H. Obaida, S. M. Elkaffas, and S. K. Guirguis, "Deep Learning Algorithms for Cyber-Bullying Detection," *IEEE Access*, vol. 12, 2024, pp. 76903–76908, doi: 10.1109/ACCESS.2024.3406595.

[2] Jain, V., Saxena, A. K., Senthil, A., Jain, A., & Jain, A., "Cyber-Bullying Detection in Social Media Platform using Machine Learning," *Proceedings of the SMART–2021, IEEE Conference*, Teerthanker Mahaveer University, Moradabad, India, Dec. 2021.

[3] Iwendi, C., Srivastava, G., Khan, S., & Maddikunta, P. K. R., "Cyberbullying detection solutions based on deep learning architectures," *Multimedia Systems*, 2020. https://doi.org/10.1007/s00530-020-00701-5

[4] Aldhyani, T. H. H., Al-Adhaileh, M. H., & Alsubari, S. N., "Cyberbullying Detection System Based on Deep Learning Algorithms," *Electronics*, vol. 11, no. 23, 2022, p. 3273. https://doi.org/10.3390/electronics11203273

[5] Shylaja, S. S., Narayanan, A., Venugopal, A., & Prasad, A., "Recurrent Neural Network Architectures with Trained Document Embeddings for Flagging Cyber-Aggressive Comments on Social Media."

[6] Fortuna, P., & Nunes, S., "A survey on automatic detection of hate speech in text," *ACM Computing Surveys*, vol. 51, no. 4, pp. 1–30, Jul. 2019, doi: 10.1145/3232676.

[7] FBI, "Hate crime statistics," 2015. [Online]. Available: https://ucr.fbi.gov/hate-crime/

[8] Burnap, P., & Williams, M. L., "Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making," *Policy & Internet*, vol. 7, no. 2, pp. 223–242, Jun. 2015.

[9] Wendling, M., *The Year That Angry Won the Internet*. London, U.K.: BBC Trending, 2015.

[10] Haidar, B., Chamoun, M., & Serhrouchni, A., "Arabic cyberbullying detection: Using deep learning," in *Proc. 7th Int. Conf. Comput. Commun. Eng. (ICCCE)*, Sep. 2018, pp. 284–289, doi: 10.1109/ICCCE.2018.8539303.

International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

[11] Malik, J. S., Qiao, H., Pang, G., & van den Hengel, A., "Deep learning for hate speech detection: A comparative study," 2022, *arXiv preprint* arXiv:2202.09517.

[12] Obaid, M. H., Guirguis, S. K., & Elkaffas, S. M., "Cyberbullying detection and severity determination model," *IEEE Access*, vol. 11, pp. 97391–97399, 2023, doi: 10.1109/ACCESS.2023.3313113.

[13] Balakrishnan, V., Khan, S., & Arabnia, H. R., "Improving cyberbullying detection using Twitter users' psychological features and machine learning," *Computers & Security*, vol. 90, Mar. 2020, Art. no. 101710, doi: 10.1016/j.cose.2019.101710.

[14] Angelis, J. D., & Perasso, G., "Cyberbullying detection through machine learning: Can technology help to prevent Internet bullying?" *International Journal of Management & Humanities*, vol. 4, no. 11, pp. 57–69, Jul. 2020, doi: 10.35940/ijmh.k1056.0741120.

[15] Prashar, S., & Bhakar, S., "Real-time cyberbullying detection," *International Journal of Engineering and Advanced Technology*, vol. 9, no. 2, pp. 5197–5201, Dec. 2019, doi: 10.35940/ijeat.b4253.129219.

[16] Yadav, J., Kumar, D., & Chauhan, D., "Cyberbullying detection using pre-trained BERT model," in *Proc. Int. Conf. Electron. Sustain. Commun. Syst. (ICESC)*, Jul. 2020, pp. 1096–1100, doi: 10.1109/ICESC48915.2020.9155700.

[17] Murshed, B. A. H., Abawajy, J., Mallappa, S., Saif, M. A. N., & Al-Ariki, H. D. E., "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.