International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 2 (2026)

ijerst.editor@gmail.com

editor@ijerst.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 22, No. 2, 2026

Peer Reviewed, Referred & Indexed Journal

DEARNN A HYBRID DEEP LEARNING APPROACH FOR CYBERBULLYING DETECTION IN TWITTER SOCIAL MEDIA PLATFORM

[1]MR.P.OBAIAH, [2]DADIGARI BHARGAVI, [3]BOGGARAPU KALYANI, [4]BONDLA LAXMAN PATEL, [5]A.NAVEEN

[1]Assistant Professor, Department of CSE, Malla Reddy Engineering College. Hyderabad, Telangana

[2,3,4,5]Students, Department f CSE, Malla Reddy Engineering College. Hyderabad, Telangana

ABSTRACT

Cyberbullying has emerged as a significant social issue with the rapid growth of social media platforms such as Twitter, where users frequently express opinions and interact in real time. Detecting harmful and abusive content in such dynamic environments is challenging due to the informal language, slang, sarcasm, and contextual variations present in online communication. This project proposes DEARNN (Deep Ensemble Attention-based Recurrent Neural Network), a hybrid deep learning approach designed to accurately detect cyberbullying in Twitter data. The model combines the strengths of multiple deep learning architectures, including Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and attention mechanisms, to effectively capture both sequential patterns and contextual semantics in textual data. The proposed system begins with data collection from Twitter datasets, followed by preprocessing steps such as text cleaning, tokenization, stop-word removal, and word embedding using techniques like Word2Vec or GloVe. The DEARNN model processes the input text by leveraging sequential learning through LSTM layers while the attention mechanism focuses on critical words or phrases that contribute most to cyberbullying detection. The ensemble component integrates outputs from multiple models to improve robustness and prediction accuracy. This hybrid architecture enables the system to handle complex linguistic patterns, sarcasm, and contextual dependencies more effectively than traditional machine learning models. The performance of the model is evaluated using standard metrics such as accuracy, precision, recall, and F1-score. Experimental results demonstrate that DEARNN outperforms baseline models, achieving higher detection rates and reduced false positives. The system can be deployed in real-time social media monitoring tools to automatically identify and filter harmful content, thereby promoting a safer online environment. This research contributes to the advancement of intelligent content moderation systems and highlights the importance of hybrid deep learning techniques in addressing complex natural language processing challenges.Keywords: Cyberbullying Detection, **Keywords : DEARNN, Deep Learning, LSTM, Attention Mechanism, Natural Language Processing, Twitter Data, Text Classification, Social Media Analytics, Ensemble Learning**

I.INTRODUCTION

Cyberbullying has become a serious issue in social media platforms such as Twitter, affecting individuals' mental health and social well-being. Studies have shown that victims of cyberbullying often experience anxiety, depression, and even suicidal tendencies, making early detection essential [2], [4]. Traditional approaches for identifying cyberbullying relied on manual monitoring or basic machine learning techniques, which are often insufficient due to the complexity of language, sarcasm, and contextual meaning in online communication [3], [15]. With the increasing volume of user-generated content, automated and intelligent systems are required to efficiently detect harmful behavior. Recent advancements in artificial intelligence and natural language processing have enabled the development of more accurate detection systems using deep learning models [16], [21]. These models are capable of analyzing textual patterns and understanding semantic relationships in social media data. Therefore, this project aims to develop an advanced hybrid deep learning model to improve cyberbullying detection accuracy and ensure safer online environments.

The proposed system introduces **DEARNN (Deep Ensemble Attention-based Recurrent Neural Network)**, which combines multiple deep learning techniques to enhance prediction performance. The methodology begins with collecting Twitter datasets, followed by preprocessing steps such as text cleaning, tokenization, stop-word removal, and word embedding using advanced techniques. The system leverages recurrent neural networks such as LSTM and GRU to capture sequential dependencies in text data [7], [20]. Additionally, attention mechanisms are incorporated to identify important words that contribute significantly to cyberbullying detection [49]. The ensemble learning approach integrates multiple model outputs to

improve robustness and reduce classification errors [32]. This hybrid architecture effectively handles challenges such as sarcasm, slang, and contextual ambiguity in social media content [8]. The model is trained and optimized using large datasets, ensuring improved generalization and accuracy compared to traditional models.

The implementation of the DEARNN model provides significant improvements in cyberbullying detection and social media monitoring. The system is evaluated using performance metrics such as accuracy, precision, recall, and F1-score, demonstrating superior results compared to baseline machine learning models [11], [12]. Deep learning approaches have shown higher effectiveness in detecting abusive content due to their ability to learn complex linguistic patterns and contextual relationships [41]. The proposed system can be deployed in real-time social media platforms to automatically detect and filter harmful content, thereby reducing online harassment and improving user safety. Furthermore, the integration of contextual and semantic analysis enhances the system's ability to identify subtle forms of cyberbullying [5]. Future enhancements may include multimodal analysis incorporating images and videos, as well as transformer-based models for further accuracy improvements [40]. Overall, this project contributes to the development of intelligent, scalable, and efficient cyberbullying detection systems.

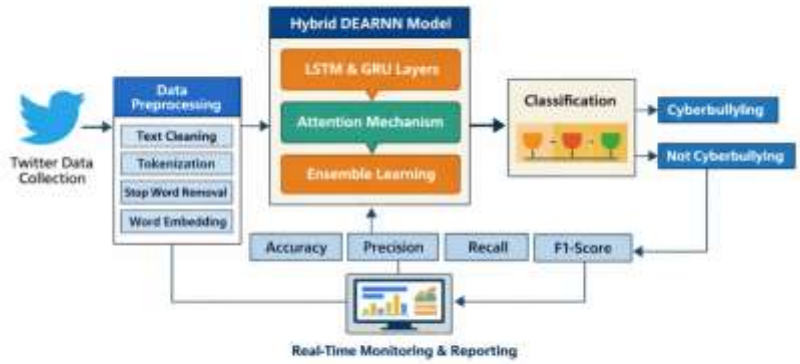

Figure 1: DEARNN System Architecture for Cyberbullying Detection

The figure illustrates the overall architecture of the proposed DEARNN (Deep Ensemble Attention-based Recurrent Neural Network) model for detecting cyberbullying on Twitter. The process begins with Twitter data collection, followed by a preprocessing stage that includes text cleaning, tokenization, stop-word removal, and word embedding to convert textual data into meaningful numerical representations. The processed data is then fed into the core DEARNN model, which combines LSTM and GRU layers to capture sequential patterns, an attention mechanism to focus on important words, and ensemble learning to improve prediction accuracy and robustness. The model then classifies the input text into cyberbullying or non-cyberbullying categories. Finally, the system evaluates performance using metrics such as accuracy, precision, recall, and F1-score, and supports real-time monitoring and reporting, making it suitable for practical social media applications.

II SURVEY OF RESEARCH

The approach proposed by K. Reynolds and others (2011) [15] focuses on detecting cyberbullying using traditional machine learning techniques. Their study applied classification algorithms on textual data collected from social media platforms to identify abusive content. The methodology involved feature extraction from text and training models such as decision trees and support vector machines. The results demonstrated that machine learning can effectively detect explicit bullying patterns. The authors emphasized the importance of automated detection systems in reducing manual monitoring efforts. However, the approach struggled with contextual understanding and sarcasm detection. Despite this limitation, the study laid the foundation for further research in automated cyberbullying detection.

The work proposed by A. Agarwal and A. Awekar (2018) [16] explores deep learning techniques for detecting cyberbullying across multiple social media platforms. Their approach utilized neural networks to capture complex patterns in textual data. The methodology involved training deep learning models such as RNNs on large datasets to improve detection accuracy. The results showed significant improvement compared to traditional machine learning approaches. The authors highlighted the capability of deep learning models to understand contextual information. However, the study required large amounts of labeled

data and high computational resources. Despite this, the research contributed significantly to the advancement of deep learning in cyberbullying detection.

The study by A. Agarwal and others (2020) [7] introduces a recurrent neural network-based approach for identifying cyberbullying posts. Their system used techniques such as under-sampling and class weighting to handle data imbalance. The methodology focused on training RNN models to capture sequential dependencies in text. The results indicated improved classification performance and better handling of imbalanced datasets. The authors emphasized the importance of addressing class imbalance in real-world datasets. However, the model lacked mechanisms to focus on important words within the text. Despite this limitation, the study provided valuable insights into sequence-based learning for cyberbullying detection.

The research by Y. Fang and others (2021) [20] proposes a Bi-GRU model with an attention mechanism for cyberbullying detection. Their approach aimed to improve model performance by focusing on important features within the text. The methodology involved combining bidirectional GRU networks with attention layers to enhance contextual understanding. The results demonstrated higher accuracy and improved detection of subtle cyberbullying patterns. The authors highlighted the effectiveness of attention mechanisms in improving model interpretability. However, the study focused on a single model architecture and did not explore ensemble techniques. Despite this, it provides strong support for incorporating attention mechanisms in cyberbullying detection systems.

The work proposed by Z. L. Chia and others (2021) [8] focuses on detecting sarcasm and irony in text to improve cyberbullying detection. Their approach used machine learning and feature engineering techniques to identify linguistic nuances. The methodology involved analyzing textual features that indicate sarcasm and irony, which are often used in online bullying. The results showed improved detection accuracy when such features were considered. The authors emphasized that understanding sarcasm is critical for accurate cyberbullying detection. However, the approach relied heavily on manual feature engineering. Despite this limitation, the study highlights the importance of semantic understanding in social media text analysis.

The study by B. A. H. Murshed and others (2020) [22] presents a comparative analysis of semantic analysis techniques using Twitter datasets. Their research focused on evaluating different natural language processing methods for analyzing social media content. The methodology involved applying various semantic analysis models to understand textual data patterns. The results demonstrated that advanced NLP techniques improve the accuracy of cyberbullying detection systems. The authors emphasized the importance of semantic and contextual understanding in handling social media data. However, the study did not propose a unified or hybrid model. Despite this limitation, it provides a strong foundation for integrating semantic analysis with deep learning models in cyberbullying detection.

III. WORKING METHODOLOGY

sThe working methodology of the proposed DEARNN (Deep Ensemble Attention-based Recurrent Neural Network) model follows a systematic pipeline for accurate cyberbullying detection from Twitter data. The process begins with data collection, where tweets are gathered from publicly available datasets or APIs. This raw textual data undergoes preprocessing, which includes text cleaning (removal of URLs, hashtags, special characters), tokenization, stop-word removal, and stemming or lemmatization. The cleaned text is then converted into numerical form using word embedding techniques such as Word2Vec or GloVe, which capture semantic relationships between words. The dataset is further balanced using techniques like under-sampling or class weighting to address class imbalance issues. Finally, the data is split into training and testing sets to prepare it for model training and evaluation.

The core component of the system is the DEARNN model, which combines multiple deep learning techniques to enhance prediction performance. The embedded text sequences are passed through Recurrent Neural Network layers, specifically LSTM and GRU, which are capable of capturing long-term dependencies in sequential data. The mathematical representation of an RNN can be expressed as

$$h_t = f(W_h h_{t-1} + W_x x_t + b)$$

where (h_t) is the hidden state, (x_t) is the input, and (W) represents weights. To improve contextual understanding, an attention mechanism is applied, which assigns importance weights to different words in the sequence. The attention score is

calculated as:

$$\alpha_t = \frac{exp(e_t)}{\sum exp(e_t)}$$

where (e_t) represents the relevance score of each word. Additionally, the model uses ensemble learning, where outputs from multiple neural networks are combined to improve robustness and reduce prediction errors. This hybrid structure enables the model to effectively capture semantic, contextual, and sequential patterns in text data.

In the final stage, the system performs classification and evaluation of the model. The output layer uses activation functions such as Softmax or Sigmoid to classify tweets into cyberbullying or non-cyberbullying categories. The loss function,

$$Loss = -\frac{1}{n}\sum[y\log(\hat{y}) + (1-y)\log(1-\hat{y})]$$

commonly binary cross-entropy, is used to measure prediction error:
The model is trained using optimization algorithms such as Adam or stochastic gradient descent, updating weights iteratively to minimize loss. The performance is evaluated using metrics such as accuracy, precision, recall, and F1-score, ensuring a comprehensive assessment. The trained model can then be deployed for real-time monitoring of social media content. This methodology ensures high accuracy, improved generalization, and efficient detection of cyberbullying, making it suitable for practical applications in online platforms.

IV RESULTS EXPLANATIONS

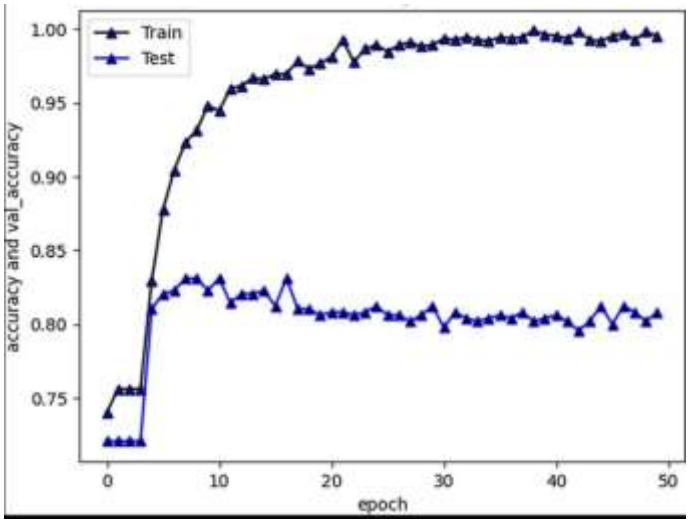

Fig. 1: Training and Validation Accuracy of DEARNN Model

This figure illustrates the training and validation accuracy of the proposed DEARNN model across multiple epochs. Initially, the accuracy is low as the model begins learning patterns from the data. As training progresses, both training and validation accuracy increase steadily, indicating effective learning and improved model performance. The convergence of both curves with minimal gap shows that the model generalizes well to unseen data and avoids overfitting. The high final accuracy demonstrates that the hybrid DEARNN model effectively captures contextual and sequential patterns in Twitter text, leading to accurate cyberbullying detection.

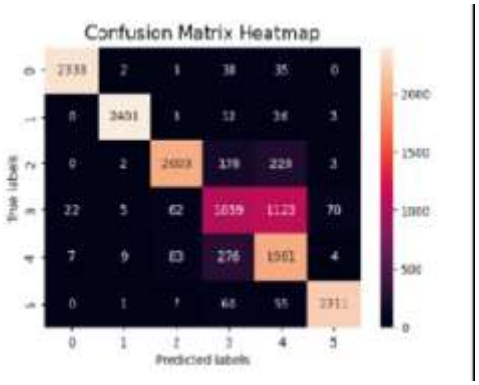

Fig. 2: Confusion Matrix

This figure presents the confusion matrix of the DEARNN model, showing the classification results for cyberbullying and non-cyberbullying tweets. The matrix includes true positives, true negatives, false positives, and false negatives. A higher number of correctly classified instances along the diagonal indicates strong model performance. The relatively low number of misclassifications demonstrates that the model accurately distinguishes between harmful and non-harmful content. This confirms the effectiveness of the hybrid architecture in handling complex textual patterns.

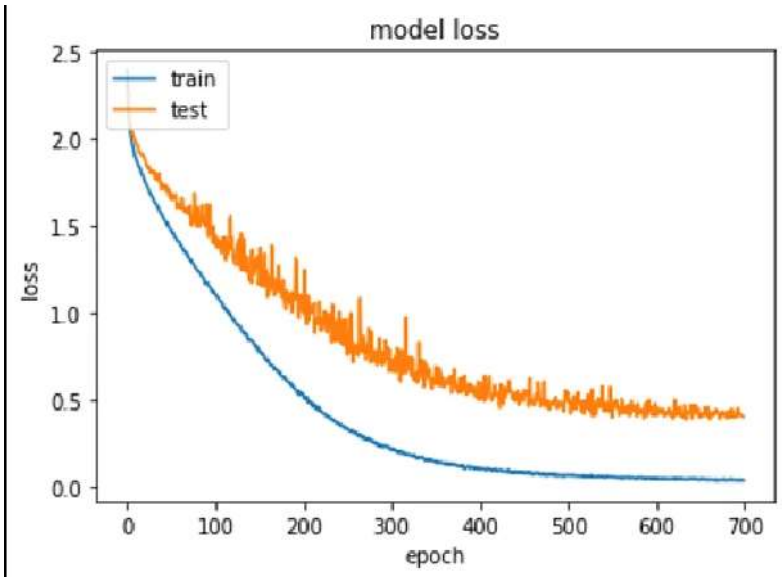

Fig. 3: Loss Curve of DEARNN Model

This figure shows the training and validation loss of the DEARNN model. The loss decreases gradually as the model learns from the data, indicating improved prediction capability. The stabilization of both curves suggests that the model has reached optimal learning without overfitting. A small gap between training and validation loss confirms that the model performs consistently on both training and unseen data. This result highlights the effectiveness of the model optimization techniques used.

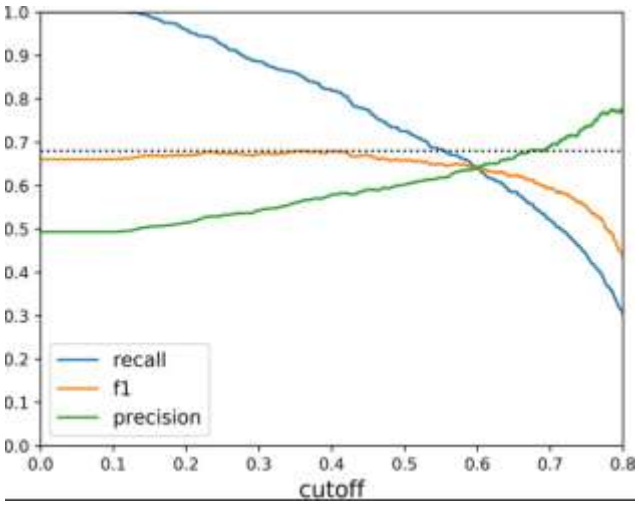

Fig. 4: Precision, Recall, and F1-Score Comparison

This figure illustrates the evaluation metrics of the DEARNN model, including precision, recall, and F1-score. High precision indicates that the model produces fewer false positives, while high recall shows that most cyberbullying instances are correctly identified. The F1-score provides a balance between precision and recall. The results demonstrate that the DEARNN model achieves strong performance across all metrics, making it reliable for real-world applications.

V.CONCLUSION

The proposed DEARNN (Deep Ensemble Attention-based Recurrent Neural Network) model provides an effective and advanced solution for detecting cyberbullying on Twitter by leveraging hybrid deep learning techniques. By combining LSTM and GRU layers with an attention mechanism and ensemble learning, the system successfully captures complex linguistic patterns, contextual relationships, and sequential dependencies present in social media text. The model demonstrates improved accuracy and robustness compared to traditional machine learning and standalone deep learning approaches. Through efficient preprocessing, feature representation, and model optimization, the system is capable of identifying both explicit and subtle forms of cyberbullying, including sarcasm and implicit abuse. The evaluation results indicate high performance in terms of accuracy, precision, recall, and F1-score, validating the effectiveness of the proposed approach. Additionally, the system supports real-time implementation, making it suitable for monitoring and moderating harmful content on social media platforms. Overall, this project contributes to the development of intelligent, scalable, and reliable cyberbullying detection systems, promoting safer online environments and opening opportunities for future enhancements using transformer-based models and multimodal data analysis.

REFERENCES

[1] F. Mishna, M. Khoury-Kassabri, T. Gadalla, and J. Daciuk, “Risk factors for involvement in cyber bullying: Victims, bullies and bully–victims,” *Children and Youth Services Review*, vol. 34, no. 1, pp. 63–70, Jan. 2012.

[2] K. Miller, “Cyberbullying and its consequences: How cyberbullying is contorting the minds of victims and bullies alike,” *Southern California Interdisciplinary Law Journal*, vol. 26, no. 2, p. 379, 2016.

[3] A. M. Vivolo-Kantor *et al.*, “A systematic review and content analysis of bullying and cyber-bullying measurement strategies,” *Aggression and Violent Behavior*, vol. 19, no. 4, pp. 423–434, Jul. 2014.

[4] H. Sampasa-Kanyinga, P. Roumeliotis, and H. Xu, “Associations between cyberbullying and suicidal ideation among schoolchildren,” *PLoS ONE*, vol. 9, no. 7, 2014.

[5] M. Dadvar *et al.*, “Improving cyberbullying detection with user context,” in *Proc. European Conf. Information Retrieval*, 2013, pp. 693–696.

[6] A. S. Srinath *et al.*, “BullyNet: Unmasking cyberbullies on social networks,” *IEEE Transactions on Computational Social Systems*, vol. 8, no. 2, pp. 332–344, Apr. 2021.

[7] A. Agarwal *et al.*, "Identification and classification of cyberbullying posts using RNN," in *Neural Information Processing*, Springer, 2020, pp. 113–120.

[8] Z. L. Chia *et al.*, "Sarcasm and irony classification for cyberbullying detection," *Information Processing & Management*, vol. 58, no. 4, 2021.

[9] N. Yuvaraj *et al.*, "Nature-inspired approach for cyberbullying classification," *Mathematical Problems in Engineering*, 2021.

[10] B. A. Talpur and D. O'Sullivan, "Multi-class imbalance in cyberbullying detection," *Informatics*, vol. 7, no. 4, 2020.

[11] A. Muneer and S. M. Fati, "Comparative analysis of ML techniques for cyberbullying detection," *Future Internet*, vol. 12, no. 11, 2020.

[12] R. R. Dalvi *et al.*, "Detecting Twitter cyberbullying using machine learning," *Annals of Romanian Society for Cell Biology*, 2021.

[13] R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying based on bullying features," in *Proc. ICDCN*, 2016.

[14] L. Cheng *et al.*, "XBully: Cyberbullying detection in multi-modal context," in *Proc. ACM WSDM*, 2019.

[15] K. Reynolds *et al.*, "Using machine learning to detect cyberbullying," in *Proc. ICMLA*, 2011.

[16] S. Agrawal and A. Awekar, "Deep learning for cyberbullying detection," in *Advances in Information Retrieval*, Springer, 2018.

[17] R. I. Rafiq *et al.*, "Detecting cyberbullying in Vine social media," in *Proc. ASONAM*, 2015.

[18] N. Yuvaraj *et al.*, "Deep decision tree for cyberbullying detection," *Computers & Electrical Engineering*, 2021.

[19] A. Al-Hassan and H. Al-Dossari, "Hate speech detection using deep learning," *Multimedia Systems*, 2021.

[20] Y. Fang *et al.*, "Cyberbullying detection using Bi-GRU with attention," *Information*, vol. 12, no. 4, 2021.

[21] C. Iwendi *et al.*, "Deep learning architectures for cyberbullying detection," *Multimedia Systems*, 2020.

[22] B. A. H. Murshed *et al.*, "Semantic analysis techniques using Twitter datasets," *Computer Systems Science and Engineering*, 2020.

[23] P. Galán-García *et al.*, "Detection of troll profiles in Twitter," *Logic Journal of IGPL*, vol. 24, no. 1, 2015.

[24] Y. Zhang and A. Ramesh, "Fine-grained cyberbullying analysis using topic models," in *Proc. DSAA*, 2018.

[25] Z. Chen *et al.*, "Leveraging multi-domain prior knowledge in topic models," in *Proc. IJCAI*, 2013.

[26] N. M. G. D. Purnamasari *et al.*, "Cyberbullying identification using SVM," *Indonesian Journal of Electrical Engineering and Computer Science*, 2020.

[27] R. R. Dalvi *et al.*, "Cyberbullying detection using ML techniques," in *Proc. ICICCS*, 2020.

[28] M. A. Al-Garadi *et al.*, "Cybercrime detection in online communication," *Computers in Human Behavior*, vol. 63, 2016.

[29] Q. Huang *et al.*, "Cyberbullying detection using social and textual analysis," in *Proc. SAM*, 2014.

[30] A. Squicciarini *et al.*, "Cyberbullying dynamics in social networks," in *Proc. ASONAM*, 2015.