RESEARCH ARTICLE OPEN ACCESS

DEA-RNN: A Hybrid Deep Learning Approach For Cyberbullying Detection in Twitter Social Media Platform

[1]Bujunuru Sudheer Kumar, [2]Dr.s. Latha

[1]Bujunuru Sudheer Kumar, MSC-CFIS, Department of computer science engineering
Dr. M.G.R educational And Research Institute, Chennai, Tamil Nadu, India
[2]Dr.s. Latha, Assistant Professor, Department of Criminology, University of Madras, Chennai, Tamil Nadu, India.

ABSTRACT

Cyberbullying (CB) has become increasingly prevalent on social media platforms, posing a significant threat to user safety across all age groups. As a result, creating a more secure online space has become a top priority. This study introduces a new hybrid deep learning framework, called DEA-RNN, specifically developed to identify cyberbullying activity on Twitter. The model integrates an Elman Recurrent Neural Network (RNN) with an improved Dolphin Echolocation Algorithm (DEA), which is employed to fine-tune crucial parameters, thereby enhancing efficiency and reducing training duration. The model was evaluated on a dataset containing 10,000 tweets and benchmarked against several established machine learning and deep learning techniques, including Bi-directional Long Short-Term Memory (Bi-LSTM), conventional RNN, Support Vector Machine (SVM), Multinomial Naive Bayes (MNB), and Random Forests (RF). Results from the experiments revealed that DEA-RNN consistently surpassed these models in accurately identifying cyberbullying content. Specifically, under the third testing scenario, the model achieved its best performance with an average accuracy of 90.45%, a precision of 89.52%, recall of 88.98%, F1-score of 89.25%, and specificity of 90.94%.

Keywords:- **—**Cyber-bullying, tweet classification, Dolphin Echolocation algorithm, Elman recurrent neural networks, short text topic modelling, cyberbullying detection, social media.

I. INTRODUCTION

Instagram has become one of the most popular online platforms for communication and social interaction across all age groups. While these platforms have revolutionized the way people connect, they also given rise to harmful activity cyberbullying. Cyberbullying is a form of psychological abuse that has a profound impact on society particularly affecting young individuals who spend significant time on social media [1].

Platforms such as Twitter and Facebook are especially vulnerable to cyberbullying due to their widespread use and the anonymity the internet provides to perpetrators. In India, for instance, 14% of all harassment cases occur on Facebook and Twitter, with 37% of these incidents involving young users. Cyberbullying can lead to severe mental health issues, including anxiety, depression, stress, and emotional distress, which, in extreme cases, may result in suicide [2].

Given these concerns, it is essential to develop effective methods for detecting cyberbullying in social media content such as posts, tweets, and comments. This article primarily focuses on detecting cyberbullying on Twitter, where the growing prevalence of online harassment highlights the need for robust identification methods and preventive strategies [3].

It is virtually infeasible to manually detect and control cyberbullying on Twitter due to the platform's vast volume of content and user interactions. Additionally, identifying cyberbullying through social media content poses significant challenges. Twitter posts are often short, filled with slang, and may include emojis or GIFs, making it difficult to accurately interpret users’ intentions and meanings based solely on text. Furthermore, cyberbullying can be subtle and hard to recognize when bullies employ tactics such as sarcasm or passive-aggressive behaviour to disguise their actions [4].

Despite these difficulties, detecting cyberbullying on social media remains an active area of research. Cyberbullying detection on Twitter has mainly relied on classifying tweets, with some studies also exploring topic modeling approaches to a lesser degree. Supervised machine learning (ML) models are commonly employed to classify tweets into bullying and non-bullying categories [5].

II. LITERATURE REVIEW

[1] Sampans-Kanyinga et al.,[6] had proved Cyberbullying and Its Psychological Impact: Cyberbullying is a growing concern due to its severe psychological effects on victims. Mishna et al. (2012) examined the risk factors for involvement in

cyberbullying, categorizing individuals into victims, bullies, and bully-victims. The study highlighted that cyberbullying often coexists with traditional bullying and has long-term psychological consequences, such as anxiety and depression (2014) found a strong correlation between cyberbullying victimization and suicidal ideation among schoolchildren, emphasizing the need for preventive strategies. Additionally, Miller (2016) explored the legal limitations in addressing cyberbullying cases, showing that current laws often fail to provide sufficient redress for victims.

[2] Dadvar&Agarwal et al.,[7]had examined machine Learning Approaches for Cyberbullying Detection: Machine learning has been widely applied in cyberbullying detection to improve accuracy and efficiency (2013) proposed an approach that enhances cyberbullying detection by incorporating user context, leading to more precise identification of bullying behaviour(2020) introduced a recurrent neural network model with under-sampling and class weighting techniques to balance the dataset, improving classification performance. Similarly, Zhao et al. (2016) developed an automatic detection system based on bullying-specific features, demonstrating its effectiveness in identifying harmful content. These studies highlight the advancements in artificial intelligence for detecting and mitigating cyberbullying.

[3] MuneerandFati et al.,[8] had proved social media platforms have become hotspots for cyberbullying, making detection a challenging task (2020) conducted a comparative analysis of various machine learning techniques for cyberbullying detection on Twitter, identifying support vector machines and deep learning as effective methods. Talpur and O' Sullivan (2020) tackled the issue of multi-class imbalance in text classification, proposing a feature engineering approach to improve cyberbullying detection. Furthermore, Cheng et al. (2019) introduced XBully, a multi-modal detection system that integrates text, images, and user behaviour to enhance cyberbullying identification. These studies emphasize the importance of robust methodologies in combating online harassment.

[4] Yuvaraj et al.,[9] had examined the Role of Feature Engineering in Cyberbullying detection: Feature engineering plays a crucial role in improving cyberbullying detection models. Chia et al. (2021) explored sarcasm and irony classification using machine learning, demonstrating that nuanced linguistic features can enhance cyberbullying detection accuracy. Yuvaraj et al. (2021) proposed a nature-inspired approach for automated classification, leveraging multimedia data for a more holistic analysis. Meanwhile, Reynolds et al. (2011) pioneered early research in using machine learning to detect cyberbullying, laying the foundation for subsequent advancements. These studies underscore the significance of feature selection and engineering in refining detection techniques.

Mishna et al., [10] had examined Risk Factors and Characteristics of Cyberbullying: Understanding the risk factors associated with cyberbullying is essential for prevention and intervention strategies. Mishna et al. (2012) identified various demographic, psychological, and behavioural characteristics that contribute to cyberbullying involvement. Their study classified individuals into victims, bullies, and bully-victims, showing how social factors and prior experiences influence their roles. Sampasa-Kanyinga et al. (2014) further explored the link between cyberbullying victimization and mental health issues, such as suicidal ideation and self-harm. These studies highlight the importance of early identification of at-risk individuals to mitigate the harmful effects of cyberbullying.

Mishna Miller et al.,[11] had proved social and Legal Responses to Cyberbullying: Legal frameworks and social interventions play a crucial role in addressing cyberbullying. Miller (2016) analysed how cyberbullying distorts the mental well-being of both victims and perpetrators, highlighting the limited effectiveness of current legal redress mechanisms. The study emphasized that online harassment laws often fail to keep pace with digital communication, leaving victims vulnerable. Furthermore, Mishna et al. (2012) discussed the role of parental guidance and school policies in preventing cyberbullying. These studies suggest that while legal measures are essential, community-based approaches and digital literacy programs are equally crucial in tackling cyberbullying.

III. Proposed Methodology

The DEA-RNN model consists of several key phases: (i) data collection, (ii) data annotation, (iii) data pre-processing and cleansing, (iv) feature extraction and selection, and (v) classification. Each of these components is outlined in detail below.

Data Collection: -The dataset comprises tweets gathered using the Twitter API streaming method, utilizing approximately 32 keywords associated with cyberbullying. These keywords, derived from psychology literature, include terms such as " idiot," " whore," " slut," " ugly," and other offensive words. Additional keywords related to hate speech, threats, and discrimination—such as “ kill,” “ terrorist,” “ racism,” and “ Islamic” —were also included. Initially, the dataset contained 435,764 tweets, with around 130,000 tweets featuring keywords associated with racism, insults, swearing, and sexism. Since only English-language tweets were required, tweets in other languages and retweets were removed as part of the filtering process. After eliminating irrelevant content, 10,000 tweets were randomly selected to form the final dataset. These pre-processing steps were carried out automatically before further refinement, which is discussed in Section III-C [12].

Pre-Processing and Data Cleansing: - The pre-processing and data cleansing stage consists of three key sub-phases. This phase is crucial for refining the raw tweet dataset and preparing it for analysis.

1.Noise Removal – This step involves eliminating unnecessary elements such as URLs, hashtags, mentions, punctuation, and symbols. Additionally, emoticons are transformed into textual representations.

2.. Out-of-Vocabulary Cleansing – This process includes spell correction, expanding acronyms, modifying removing unnecessary character elongations (e.g., repeated letters). Tweet Transformations – To standardize the text, all tweets are converted to lowercase, followed by stemming, tokenization (word segmentation), and the removal of stop words. These pre-processing steps enhance the dataset’ s quality, making it more suitable for feature extraction and improving classification accuracy. Figure 2 illustrates the complete pre-processing and data cleansing workflow [13].

Data Annotation: -Once the dataset was refined, the selected 10,000 tweets were manually labelled to classify them as either “ 0” (non-cyberbullying) or “ 1” (cyberbullying). A team of three human annotators completed this process over one and a half months, following guidelines outlined in prior research. The classification was based on factors such as insults, name-calling, mockery, threats, character attacks, verbal abuse, and physical appearance-related remarks. Initially, two annotators independently classified the tweets, achieving an agreement rate of approximately 91%. Any discrepancies were reviewed and resolved by a third annotator to finalize the dataset. The resulting labelled dataset contained 6,508 non-cyberbullying tweets (65%) and 3,492 cyberbullying tweets (35%), highlighting an imbalance between the two categories. To address this imbalance, a balancing technique was applied. The Synthetic Minority Oversampling Technique (SMOTE) was used to oversample the cyberbullying tweets, ensuring a more balanced dataset. This approach involved generating synthetic examples of cyberbullying tweets to match the number of non-cyberbullying instances, improving the model’ s ability to detect cyberbullying effectively [14].

Feature Extraction and Selection: - Future extraction from the Twitter dataset is performed using Natural Language Processing (NLP) techniques such as Word2Vec and Term Frequency-Inverse Document Frequency (TF-IDF). Key features include nouns, pronouns, and adjectives, while adverbs and verbs offer supplementary context. Additionally, extracting Part-of-Speech (POS) tags, function words, and content words can enhance classification performance**.** For feature selection, various methods have been explored in prior research. To identify cyberbullying instances effectively, the Information Gain (IG) method is applied to select the most relevant features. These selected feature subsets are then used as input for the DEA-RNN classifier to improve detection accuracy [15].

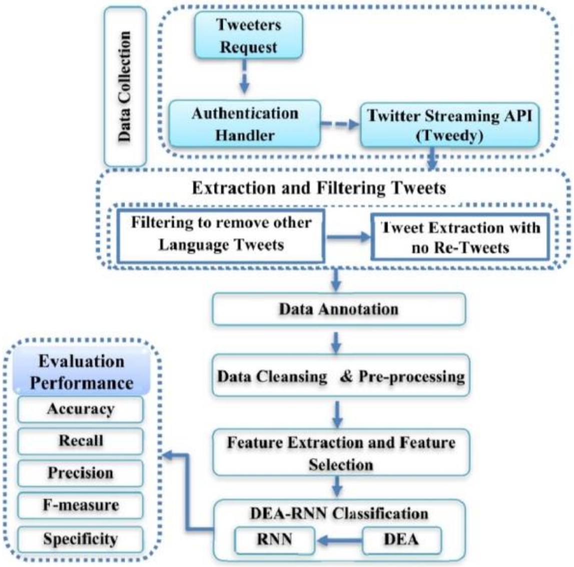

Data Collection
Tweeters Request
Authentication Handler
Twitter Streaming API (Tweedy)
Extraction and Filtering Tweets
Filtering to remove other Language Tweets
Tweet Extraction with no Re-Tweets
Data Annotation
Data Cleansing & Pre-processing
Feature Extraction and Feature Selection
DEA-RNN Classification
RNN
DEA
Evaluation Performance
Accuracy
Recall
Precision
F-measure
Specificity

Figure 1

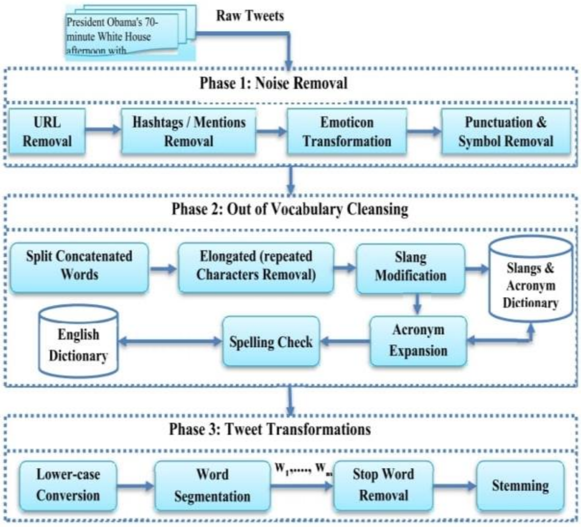

President Obama's 70-minute White House afternoon with...
Raw Tweets
Phase 1: Noise Removal
URL Removal
Hashtags / Mentions Removal
Emoticon Transformation
Punctuation & Symbol Removal
Phase 2: Out of Vocabulary Cleansing
Split Concatenated Words
Elongated (repeated Characters Removal)
Slang Modification
Slangs & Acronym Dictionary
English Dictionary
Spelling Check
Acronym Expansion
Phase 3: Tweet Transformations
Lower-case Conversion
Word Segmentation
$W_1, ..., W_n$
Stop Word Removal
Stemming

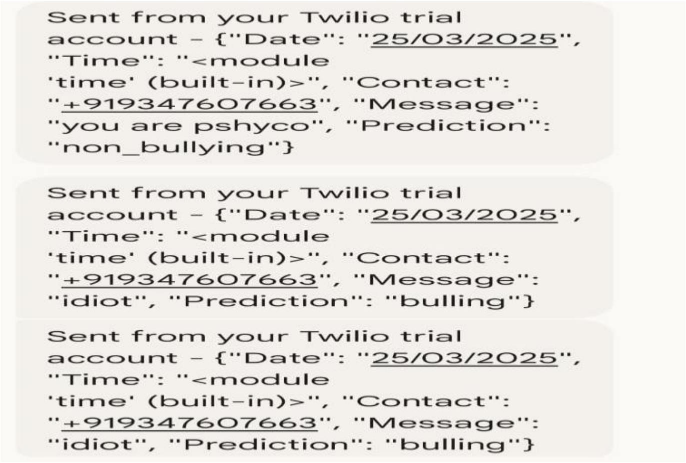
Figure 2

IV.RESULT AND DISCUSSION: -

The Performance Improvement Rate (PIR) is utilized to assess how effectively the proposed model performs by evaluating key indicators such as specificity, F1-score, precision, recall, and accuracy. PIR is determined by comparing the proposed model's overall outcomes with five baseline models—comprising two deep learning approaches and three traditional machine learning algorithms. In Scenario 2, the proposed system achieves accuracy gains of 3.69%, 6.91%, 10.04%, 12%, and 22.69% over Bi-LSTM, RNN, SVM, RF, and MNB, respectively. For Scenario 3, the accuracy increases are 1.71%, 3.3%, 5.24%, 7%, and 8.19% when evaluated against the same benchmarks. With regard to precision, the model demonstrates improvements of 4.14%, 6.93%, 10.42%, 8.06%, and 11.24% in Scenario 2 over Bi-LSTM, RNN, SVM, RF, and MNB, respectively. Scenario 3 shows corresponding precision enhancements of 1.62%, 2.9%, 5.27%, 5.65%, and 9.51%. In terms of recall, Scenario 2 yields performance gains of 4.33%, 7.34%, 10.03%, 17.24%, and 6.48%, while in Scenario 3, the recall improvement rates are 1.46%, 3.08%, 6.26%, 6.48%, and 10.09%, all relative to the same baseline models [16].

TABLE 1. The details of twitter dataset versions.

Dataset	Total number of tweets	Number of Cyberbullying Tweets	# of Non-Cyberbullying Tweets
Original Twitter	10,000	3,492	6,508
Oversampled Twitter	13016	6,508	6,508



```
Sent from your Twilio trial
account - {"Date": "25/03/2025",
"Time": "<module
'time' (built-in)>", "Contact":
"+919347607663", "Message":
"you are pshyco", "Prediction":
"non_bullying"}

Sent from your Twilio trial
account - {"Date": "25/03/2025",
"Time": "<module
'time' (built-in)>", "Contact":
"+919347607663", "Message":
"idiot", "Prediction": "bulling"}

Sent from your Twilio trial
account - {"Date": "25/03/2025",
"Time": "<module
'time' (built-in)>", "Contact":
"+919347607663", "Message":
"idiot", "Prediction": "bulling"}
```

output

V.CONCLUSIONS

This study introduced an effective tweet classification model designed to enhance topic modelling techniques for detecting cyberbullying incidents. The DEA-RNN model was developed by integrating DEA optimization with an Elman-type Recurrent Neural Network (RNN) to optimize parameter tuning. The model' s performance was evaluated against existing methods, including Bi-LSTM, RNN, SVM, RF, and MNB, using a newly curated Twitter dataset containing cyberbullying-related keywords. Experimental results demonstrated that DEA-RNN outperformed these baseline models across various metrics, including accuracy, recall, F-measure, precision, and specificity. These findings highlight the effectiveness of DEA in improving RNN performance. However, despite its high accuracy, the model' s feature compatibility diminishes when the dataset size significantly exceeds the initial input. This investigation was confined to Twitter, which restricts the analysis to just one social media platform. Future studies should extend cyberbullying detection to other platforms such as Instagram, YouTube, Facebook, and Flickr to analyse broader trends. Additionally, this study only examined textual content, without considering user behaviour—an aspect that could be explored in future research [17].

REFERENCES

[1] Mishna, M. Khoury-Kassabri, T. Gadalla, and J. Daciuk, ‘ ‘ Risk factorsfor involvement in cyber bullying: Victims, bullies and bully– victims, ‘ Children Youth Services Rev., vol. 34, no. 1, pp. 63– 70, Jan. 2012, Doi: 10.1016/j.childyouth.2011.08.032.

[2] K. Miller, ‘ ‘ Cyberbullying and its consequences: How cyberbullying is contorting the minds of victims and bullies alike, and the law’ s limited available redress,’ ’ Southern California Interdisciplinary. Law J., vol. 26, no. 2, p. 379, 2016.

[3] A. M. Vivolo-Kantor, B. N. Martell, K. M. Holland, and R. Westby, ‘ A systematic review and content analysis of bullying and cyber-bullying measurement strategies,’ ’ Aggression Violent Behave., vol. 19, no. 4, pp. 423– 434, Jul. 2014, Doi: 10.1016/j.avb.2014.06.008.

[4] H. Sampasa-Kanyinga, P. Roumeliotis, and H. Xu, ‘ ‘ Associations between cyberbullying and school bullying victimization and suicidal ideation, plans and attempts among Canadian schoolchildren,’ ’ Plops ONE, vol. 9, no. 7, Jul. 2014, Art. no. e102145, Doi: 10.1371/journal.pone.0102145.

[5] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong, ‘ ‘ Improving cyberbullying detection with user context,’ ’ in Proc. Eur. Conf. Inf. Retro., in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 7814,2013, pp. 693– 696.

[6] A. S. Srinath, H. Johnson, G. G. Dagher, and M. Long, ‘ ‘ Bully Net: Unmasking cyberbullies on social networks,’ ’ IEEE Trans.Computat. Social Syst., vol. 8, no. 2, pp. 332– 344, Apr. 2021, doi:10.1109/TCSS.2021.3049232.

[7] A. Agarwal, A. S. Chivukula, M. H. Bhuyan, T. Jan, B. Narayan, and M. Prasad, ‘ ‘ Identification and classification of cyberbullying posts: A recurrent neural network approach using under-sampling and class weighting,’ ’ in Neural Information Processing (Communications in Computer and Information Science), vol. 1333, H. Yang, K. Pasupa,A. C.-S. Leung, J. T. Kwok, J. H. Chan, and I. King, Eds. Cham, Switzerland: Springer, 2020, pp. 113– 120.

[8] Z. L. Chia, M. Ptaszynski, F. Masui, G. Leliwa, and M. Wroczynski,‘ ‘ Machine learning and feature engineering-based study into sarcasm and irony classification with application to cyberbullying detections. Process. Manage., vol. 58, no. 4, Jul. 2021, Art. no. 102600, Doi: 10.1016/j.ipm.2021.102600.

[9] N. Yuvaraj, K. Srihari, G. Dhiman, K. Somasundaram, A. Sharma, S. Rajeskannan, M. Soni, G. S. Gaba, M. A. AlZain, and M. Masud, ‘ ‘ Nature-inspired-based approach for automated cyberbullying classification on multimedia social networking,’ ’ Math. Problems Eng., vol. 2021, pp. 1– 12, Feb. 2021, Doi: 10.1155/2021/6644652.

[10] B. A. Talpur and D. O’ Sullivan, ‘ ‘ multi-class imbalance in text classification: A feature engineering approach to detect cyberbullying in Twitter,’ ’ Informatics, vol. 7, no. 4, p. 52, Nov. 2020, Doi: 10.3390/informatics7040052.

[11] A. Muneer and S. M. Fati, ‘ ‘ A comparative analysis of machine learning techniques for cyberbullying detection on Twitter,’ ’ Future. Internet, vol. 12, no. 11, pp. 1– 21, 2020, Doi: 10.3390/fi12110187.

[12] R. R. Dalvi, S. B. Chavan, and A. Halbe, ‘ ‘ Detecting a Twitter cyberbullying using machine learning,’ ’ Ann. Romanian Soc. Cell Biol., vol. 25, no. 4, pp. 16307– 16315, 2021.

[13] R. Zhao, A. Zhou, and K. Mao, ‘ ‘ Automatic detection of cyberbullying on social networks based on bullying features,’ ’ in Proc. 17th Int. Conf. Disturb. Compute. Newt., Jan. 2016, pp. 1– 6, Doi: 10.1145/2833312.2849567.

[14] L. Cheng, J. Li, Y. N. Silva, D. L. Hall, and H. Liu, ‘ ‘ XBully: Cyberbullying detection within a multi-modal context,’ ’ in Proc. 12th ACM Int. Conf. Web Search Data Mining, Jan. 2019, pp. 339– 347, doi:10.1145/3289600.3291037.

[15] K. Reynolds, A. Kontostathis, and L. Edwards, ‘ ‘ Using machine learning to detect cyberbullying,’ ’ in Proc. 10th Int. Conf. Mach. Learn. Appl. Workshops (ICMLA), vol. 2, Dec. 2011, pp. 241– 244, doi:10.1109/ICMLA.2011.152.

[16] R I Rafiq, H. Hosseinmardi, R. Han, Q. Lv, S. Mishra, and S. A. Mattson, ‘ Careful what you share in six seconds: Detecting cyberbullying instances in vine,’ ’ in Proc. IEEE/ACM Int. Conf. Adv. Social Newt. Anal. Mining (ASONAM), Aug. 2015, pp. 617– 622, Doi: 10.1145/2808797.2809381.

[17] N. Yuvaraj, V. Chang, B. Gobinathan, A. Pinagapani, S. Kannan,G. Dhiman, and A. R. Rajan, ‘ ‘ Automatic detection of cyberbullying using -feature based artificial intelligence with deep decision tree classic-,’ ’ Compute. Electra. Eng., vol. 92, Jun. 2021, Art. No. 107186, Doi: 10.1016/j.compeleceng.2021.107186.