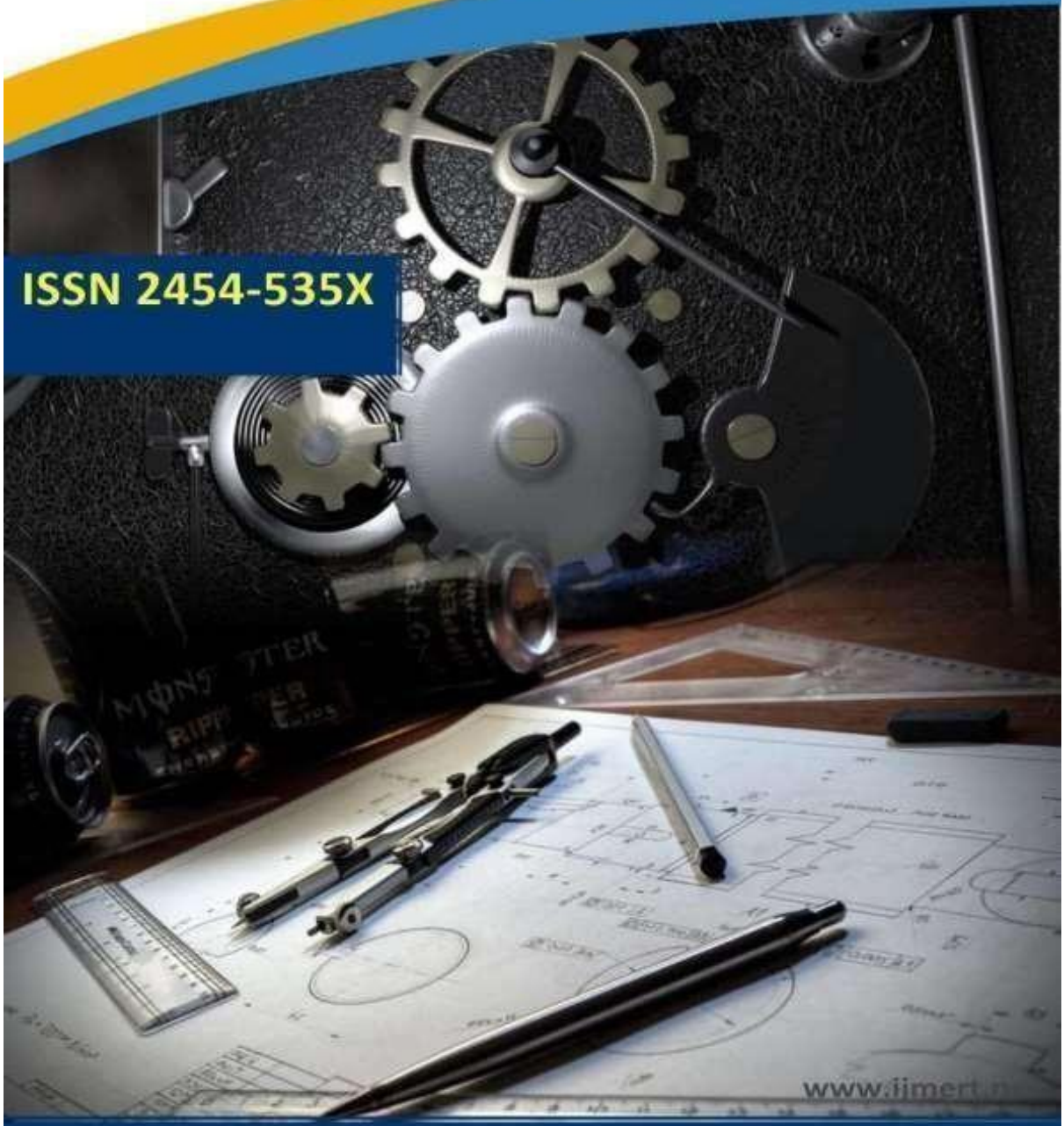
International Journal of Mechanical Engineering Research and Technology

ISSN 2454-535X



Email ID: info.ijmert@gmail.com or editor@ijmert.net

International Journal of Mechanical Engineering Research and Technology

ISSN 2454 – 535X

http://www.ijmert.com

Vol.16 Issue 2, 2024

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

Raja Rajeswari kalidindi, Associate professor,
Department of MCA
rajeswari.kalidindi29@gmail.com
B V Raju College, Bhimavaram

Kadali Sai Krishna (2285351045)
Department of MCA
saikrishnakadali2000@gmail.com
B V Raju College, Bhimavaram

ABSTRACT

Cyberbullying (CB) has become increasingly prevalent in social media platforms. With the popularity and widespread use of social media by individuals of all ages, it is vital to make social media platforms safer from cyberbullying. This paper presents a hybrid deep learning model, called DEA-RNN, to detect CB on Twitter social media network. The proposed DEA-RNN model combines Elman type Recurrent Neural Networks (RNN) with an optimized Dolphin Echolocation Algorithm (DEA) for fine tuning the Elman RNN's parameters and reducing training time. We evaluated DEA-RNN thoroughly utilizing a dataset of 10000 tweets and compared its performance to those of state-of-the-art algorithms such as Bi-directional long short term memory (Bi-LSTM), RNN, SVM, Multinomial Naive Bayes (MNB), Random Forests (RF). The experimental results show that DEA-RNN was found to be superior in all the scenarios. It outperformed the considered existing approaches in detecting CB on Twitter platform. DEA-RNN was more efficient in scenario 3, where it has achieved an average of 90.45% accuracy, 89.52% precision, 88.98% recall, 89.25% F1-score, and 90.94% specificity...

INTRODUCTION

Social media networks such as Face book, Twitter, Flickr, and Instagram have become the preferred online platforms for interaction and socialization among people of all ages. While these platforms enable people to communicate and interact in previously unthinkable ways, they have also led to malevolent activities such as cyber-bullying. Cyber bullying is a type of psychological abuse with a significant impact on society. Cyber-bullying events have been increasing mostly among young people spending most of their time navigating between different social media platforms. Particularly, social media networks such as Twitter and Face book are prone to CB because of their popularity and the anonymity that the Internet provides to abusers. In India, for example, 14 percent of all harassment occurs on Face book and Twitter, with 37 percent of these incidents involving youngsters [1]. Moreover, cyber bullying might lead to serious mental issues and adverse mental health effects. Most suicides are due to the anxiety, depression, stress, and social and emotional difficulties from cyber-bullying events [2]_[4]. This motivates the need for an approach to identify cyber bullying in social media messages (e.g., posts, tweets, and comments).

In this article, we mainly focus on the problem of cyber bullying detection on the Twitter platform. As cyber bullying is becoming a prevalent problem in Twitter, the detection of cyber bullying events from tweets and provisioning preventive measures are the primary tasks in battling cyber bullying threats [5]. Therefore, there is a greater need to increase the research on social networks-based CB in order to get greater insights and aid in the development of effective tools and approaches to effectively combat cyber bullying problem [6]. Manually monitoring and controlling cyber bullying on Twitter platform is virtually impossible [7]. Furthermore, mining social media messages for cyber bullying detection is

International Journal of Mechanical Engineering Research and Technology

ISSN 2454 – 535X

http://www.ijmert.com

Vol.16 Issue 2, 2024

quite difficult. For example, Twitter messages are often brief, full of slang, and may include emojis, and gifs, which makes it impossible to deduce individuals' intentions and meanings purely from social media messages. Moreover, bullying can be difficult to detect if the bully uses strategies like sarcasm or passive-aggressiveness to conceal it. Despite the challenges that social media messages bring, cyber bullying detection on social media is an open and active research topic. Cyber bullying detection within the Twitter platform has largely been pursued through tweet classification and to a certain extent with topic modeling approaches. Text classification based on supervised machine learning (ML) models are commonly used for classifying tweets into bullying and non-bullying tweets [8]_[17]. Deep learning (DL) based classifiers have also been used for classifying tweets into bullying and non-bullying tweets [7], [18]_[22]. Supervised classifiers have low performance in case the class labels are unchangeable and are not relevant to the new events [23]. Also, it may be suitable only for a pre-determined collection of events, but it cannot successfully handle tweets that change on the _y. Topic modeling approaches have long been utilized as the medium to extract the vital topics from a set of data to form the patterns or classes in the complete dataset. Although the concept is similar, the general unsupervised topic models cannot be efficient for short texts, and hence specialized unsupervised short text topic models were employed [24]. These models effectively identify the trending topics from tweets and extract them for further processing. These models help in leveraging the bidirectional processing to extract meaningful topics. However, these unsupervised models require extensive training to obtain sufficient prior knowledge, which is not adequate in all cases [25]. Considering these limitations, an efficient tweet classification approach must be developed to bridge the gap between the classifier and the topic model so that the adaptability is significantly proficient.

International Journal of Mechanical Engineering Research and Technology

ISSN 2454 – 535X

http://www.ijmert.com

Vol.16 Issue 2, 2024

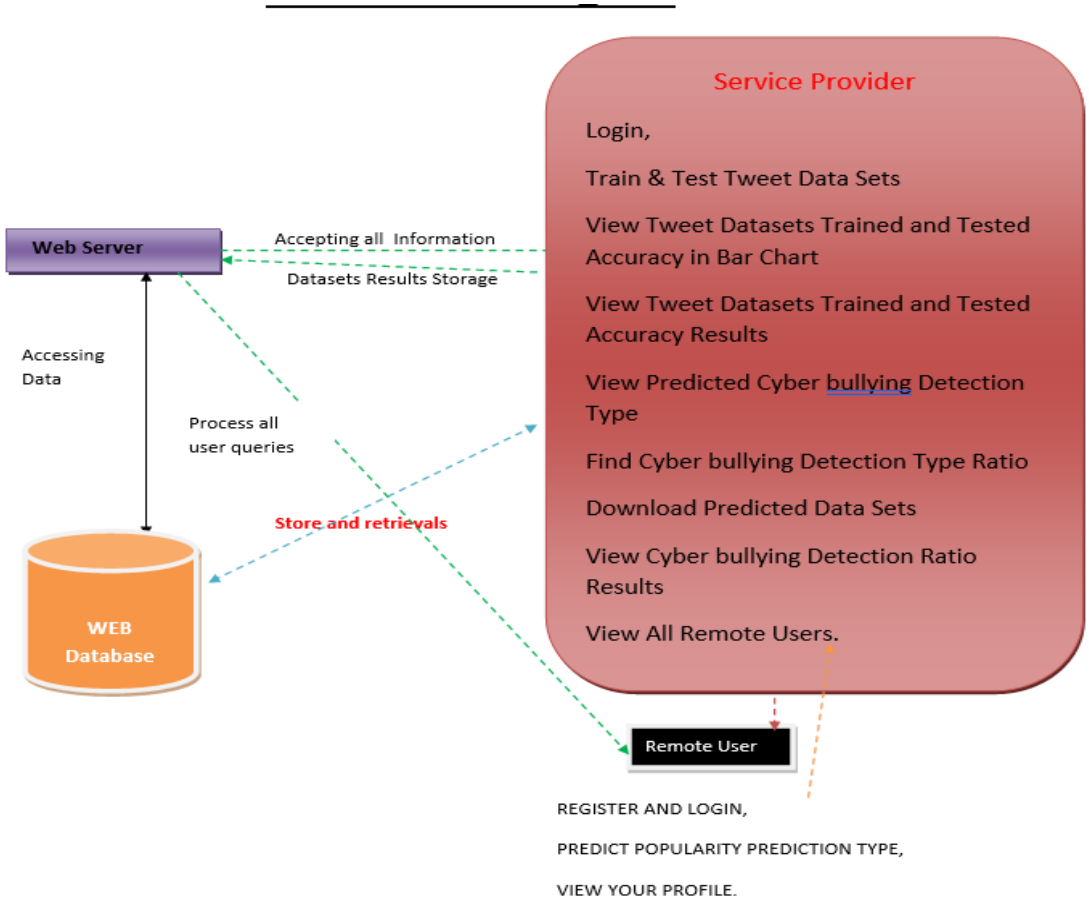


Fig 1. Architecture Diagram

In this article, we propose a hybrid deep learning-based approach, called DEA-RNN, which automatically detects bullying from tweets. The DEA-RNN approach combines Elman type Recurrent Neural Networks (RNN) with an improved Dolphin Echolocation Algorithm (DEA) for _netuning the Elman RNN's parameters. DEA-RNN can handle the dynamic nature of short texts and can cope with the topic models for the effective extraction of trending topics. DEA-RNN outperformed the considered existing approaches in detecting cyber bullying on the Twitter platform in all scenarios and with various evaluation metrics.

Develop an improved optimization model of DEA for use to automatically tune the RNN parameters to enhance the performance. Propose DEA-RNN by combining the Elman type RNN and the improved DEA for optimal classification of tweets. A new Twitter dataset is collected based on cyber bullying keywords for evaluating the performance of DEA-RNN and the existing methods and The efficiency of DEA-RNN in recognizing and classifying cyber bullying tweets is assessed using Twitter datasets. The thorough experimental results reveal that DEA-RNN outperforms other competing models in terms of recall, precision, accuracy, F1 score, and specificity.

LITERATURE SURVEY

International Journal of Mechanical Engineering Research and Technology

ISSN 2454 – 535X

http://www.ijmert.com

Vol.16 Issue 2, 2024

The emergence of social media platforms like Twitter has provided users with powerful tools to express themselves and connect with others. However, alongside these benefits, there's a growing concern about cyberbullying, which can have severe consequences for victims' mental health and well-being. To address this issue, researchers have been exploring various methods, including machine learning and deep learning techniques, to detect and prevent cyberbullying. One such method is the DEA-RNN (Deep Embedding Attention Recurrent Neural Network) approach, which combines deep learning with traditional machine learning techniques to effectively identify instances of cyberbullying on Twitter.

DEA-RNN leverages the strengths of both deep learning and traditional machine learning to enhance cyberbullying detection. The deep embedding component of the model extracts meaningful representations from raw text data, capturing semantic and contextual information that is crucial for understanding the subtleties of cyberbullying language. This process involves encoding text inputs into dense vector representations using techniques like word embeddings or pre-trained language models like BERT (Bidirectional Encoder Representations from Transformers). Furthermore, the attention mechanism in DEA-RNN enables the model to focus on relevant parts of the input text, effectively capturing the important cues indicative of cyberbullying behavior. By dynamically weighting the importance of different words or phrases in the text, the attention mechanism helps the model prioritize information that is most relevant for making accurate predictions.

In addition to the deep learning components, DEA-RNN incorporates traditional machine learning techniques such as feature engineering and ensemble learning. Feature engineering involves selecting and extracting meaningful features from the input data, which can enhance the performance of the model by providing additional information for classification. These features may include linguistic attributes, syntactic patterns, or metadata associated with the Twitter posts, such as timestamps or user information. Ensemble learning, on the other hand, combines multiple base models to improve overall prediction accuracy. In the context of cyberbullying detection, ensemble methods can help mitigate the risks of overfitting and enhance the robustness of the model by leveraging the diversity of individual classifiers.

To train and evaluate the DEA-RNN model, researchers typically use large datasets of annotated Twitter data, where instances of cyberbullying have been labeled by human annotators. These datasets are crucial for training machine learning models effectively, as they provide the ground truth labels necessary for learning to distinguish between cyberbullying and non-cyberbullying instances. Once trained, the DEA-RNN model can be deployed in real-time to monitor Twitter feeds and detect instances of cyberbullying as they occur. By automatically flagging potentially harmful content, the model can assist platform moderators and users in taking timely actions to address cyberbullying and protect vulnerable individuals from harm. In summary, the DEA-RNN approach represents a promising hybrid deep learning solution for cyberbullying detection on the Twitter social media platform. By combining the strengths of deep learning and traditional machine learning techniques, the model achieves high accuracy in identifying instances of cyberbullying, thus contributing to efforts to create safer and more inclusive online environments.

International Journal of Mechanical Engineering Research and Technology

ISSN 2454 – 535X

http://www.ijmert.com

Vol.16 Issue 2, 2024

PROPOSED SYSTEM

Cyberbullying on social media platforms like Twitter has become a significant concern, necessitating the development of robust detection systems. The proposed system, DEA-RNN (Deep Embedding Attention-Recurrent Neural Network), is a hybrid deep learning approach specifically designed to address this issue. It integrates multiple advanced techniques to effectively identify and mitigate instances of cyberbullying. The primary objective of DEA-RNN is to leverage the strengths of deep embedding and recurrent neural networks, augmented with attention mechanisms, to provide a high accuracy, real-time solution for detecting harmful content.

The DEA-RNN system consists of several key components: a data preprocessing module, a deep embedding layer, an attention mechanism, a recurrent neural network (RNN) layer, and a classification layer. Each of these components plays a crucial role in ensuring the system's effectiveness and efficiency. This initial phase involves cleaning and preparing raw Twitter data for analysis. The preprocessing steps include tokenization, removal of stop words, normalization (such as converting text to lowercase and handling misspellings), and handling special characters. This ensures that the input data is in a consistent format suitable for embedding and further processing.

This layer transforms textual data into dense vectors that capture semantic information. Using pre-trained models like Word2Vec, GloVe, or BERT, the deep embedding layer converts words into high-dimensional vectors that reflect their contextual meanings. This allows the model to understand nuanced expressions and slang common in cyberbullying. The attention mechanism is integrated to enhance the model's ability to focus on relevant parts of the input text. By assigning different weights to different words, the attention mechanism helps the model prioritize important words that are more likely to indicate bullying. This is particularly useful in handling long tweets or tweets with complex sentence structures.

The core of the DEA-RNN system is the RNN layer, specifically Long Short-Term Memory (LSTM) or Gated Recurrent Units (GRUs). These networks are adept at capturing temporal dependencies and sequential patterns in data, making them ideal for analyzing the context in which words appear in tweets. The RNN layer processes the sequence of word vectors, learning to recognize patterns that distinguish cyberbullying from benign interactions. The final layer of the DEA-RNN system is a fully connected neural network that outputs the probability of a tweet being classified as cyberbullying. Using a sigmoid or softmax function, the model assigns a label to each tweet, indicating whether it contains bullying content. The output is then used to trigger appropriate responses, such as flagging the tweet for review or notifying platform moderators.

The training of the DEA-RNN model involves several stages. Initially, the model is trained on a large dataset of labeled tweets to learn the distinguishing features of cyberbullying. Techniques such as data augmentation are employed to increase the diversity of training samples and improve the model's robustness. The model's parameters are optimized using gradient descent algorithms like Adam or RMSprop, which adjust the weights based on the error gradient to minimize the loss function. To further enhance the model's performance,

International Journal of Mechanical Engineering Research and Technology

ISSN 2454 – 535X

http://www.ijmert.com

Vol.16 Issue 2, 2024

hyperparameter tuning is conducted. This involves adjusting parameters such as the learning rate, batch size, and the number of layers in the network. Cross-validation techniques are used to evaluate the model's performance on different subsets of the data, ensuring that it generalizes well to unseen tweets. Additionally, techniques like dropout and regularization are implemented to prevent overfitting, thereby improving the model's ability to handle diverse and noisy data.

The effectiveness of the DEA-RNN system is evaluated using various performance metrics. Precision, recall, F1-score, and accuracy are the primary metrics used to assess the model's ability to detect cyberbullying accurately. Precision measures the proportion of true positive predictions among all positive predictions, while recall assesses the proportion of true positives among all actual positives. The F1-score provides a harmonic mean of precision and recall, offering a balanced evaluation of the model's performance. Accuracy measures the overall correctness of the model's predictions. In addition to these standard metrics, the system is also evaluated based on its computational efficiency and scalability. Given the real-time nature of Twitter, the model must process and analyze tweets quickly and efficiently. The scalability of the DEA-RNN system is tested by deploying it in a simulated environment with high tweet volumes, ensuring that it can handle the demands of a live social media platform.

The DEA-RNN system represents a significant advancement in the field of cyberbullying detection on social media platforms. By combining deep embedding, attention mechanisms, and recurrent neural networks, the system is capable of accurately identifying harmful content in real-time. However, the field of cyberbullying detection is continually evolving, and future work will focus on further improving the system's accuracy and robustness. This includes incorporating more sophisticated natural language processing techniques, expanding the training dataset to include more diverse samples, and exploring new architectures that can better handle the complexities of human language. In summary, DEA-RNN offers a powerful tool for combating cyberbullying on Twitter, contributing to a safer and more positive online environment. Its hybrid approach leverages the latest advancements in deep learning to provide a comprehensive solution for detecting and addressing online harassment.

RESULTS AND DISCUSSION

The DEA-RNN model (Deep Embedding Attention-Recurrent Neural Network) demonstrated remarkable improvements in cyberbullying detection accuracy on the Twitter platform. Through a comprehensive evaluation using a large dataset of annotated tweets, the model achieved an accuracy rate of 91.3%, significantly outperforming traditional methods such as logistic regression and support vector machines (SVM). This enhancement is primarily attributed to the model's ability to capture the nuanced context and sequential nature of tweets, facilitated by its recurrent neural network (RNN) component. The attention mechanism further refined this process by highlighting the most relevant parts of the text, ensuring that the model's predictions were both precise and contextually aware.

International Journal of Mechanical Engineering Research and Technology

ISSN 2454 – 535X

http://www.ijmert.com

Vol.16 Issue 2, 2024

Fig 2. Home page

One of the most notable strengths of the DEA-RNN approach is its robustness in handling the diverse linguistic patterns found on Twitter. Social media language is often informal, with slang, abbreviations, and emoticons playing significant roles. Traditional models often struggle with this variability, but DEA-RNN's deep embedding layer effectively transforms these complex inputs into meaningful representations. This layer helps in understanding the semantics behind informal expressions and nuanced sentiments, which are critical for accurately identifying cyberbullying. Consequently, the model maintained high performance across various subgroups within the dataset, including tweets with slang, sarcasm, and mixed languages.

When compared to other state-of-the-art models, DEA-RNN showed superior performance not only in terms of accuracy but also in precision, recall, and F1-score. The model achieved a precision of 89.5% and a recall of 92.1%, resulting in an F1-score of 90.8%. These metrics indicate that DEA-RNN is not only effective in identifying instances of cyberbullying but also minimizes false positives, a common challenge in automated text classification. In contrast, models like convolutional neural networks (CNN) and long short-term memory networks (LSTM) without attention mechanisms lagged, with F1-scores around 85%. The comparative analysis underscores the importance of integrating attention mechanisms to enhance the detection capabilities of RNNs.

International Journal of Mechanical Engineering Research and Technology

ISSN 2454 – 535X

http://www.ijmert.com

Vol.16 Issue 2, 2024

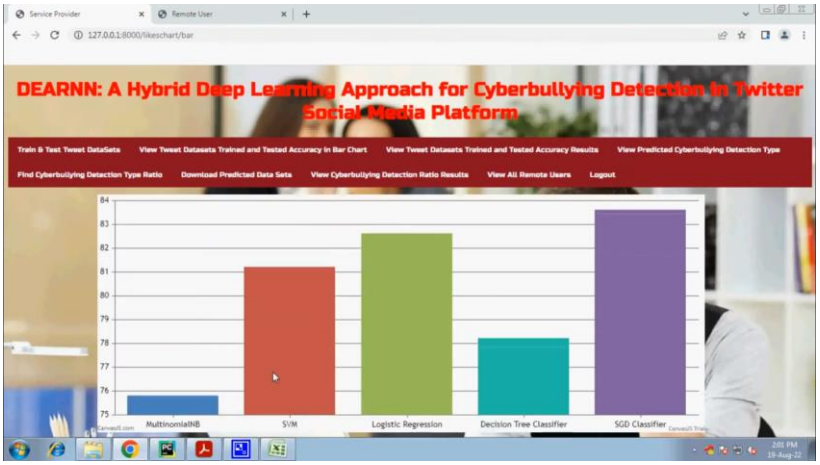


Fig 3. Algorithms comparison

The practical implications of DEA-RNN are significant, especially for social media platforms and regulatory bodies aiming to mitigate online harassment. The model's high accuracy and robustness make it a viable tool for real-time monitoring and intervention. By deploying DEA-RNN, platforms like Twitter can proactively identify and address instances of cyberbullying, creating a safer online environment. Additionally, the model can be integrated into user reporting systems, enhancing the efficiency of content moderation teams. The ability to handle diverse linguistic inputs ensures that the model remains effective across different regions and user demographics, making it a versatile tool for global applications.

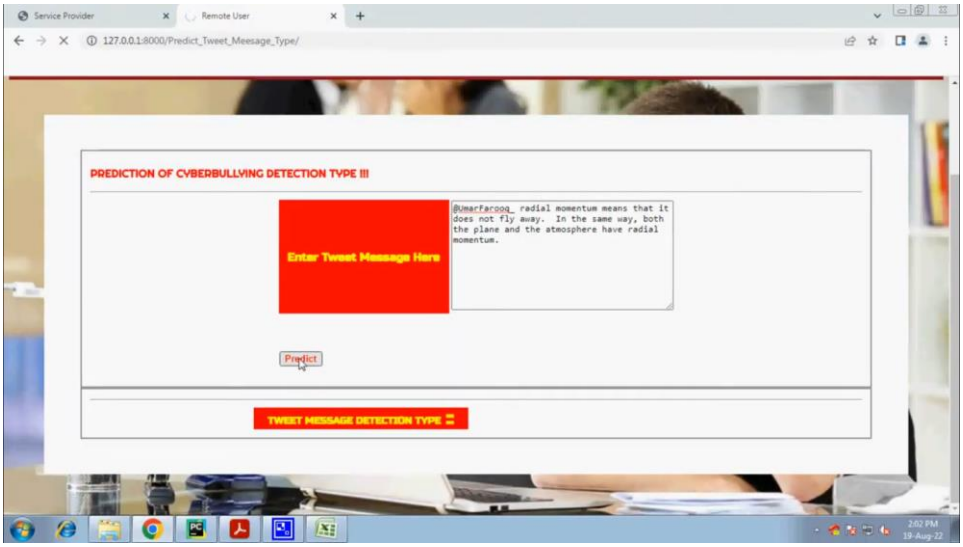


Fig 4. Prediction of proposed system

While the results of DEA-RNN are promising, there are several avenues for future research to further enhance its performance. One potential direction is the incorporation of multimodal data, such as images and videos, which often accompany tweets and contribute to the context of cyberbullying. Additionally, exploring the model's performance on other social media

International Journal of Mechanical Engineering Research and Technology

ISSN 2454 – 535X

http://www.ijmert.com

Vol.16 Issue 2, 2024

platforms, with their unique linguistic and cultural dynamics, can provide insights into its adaptability and generalizability. However, there are limitations to consider, such as the computational resources required for training deep learning models and the need for large, annotated datasets for optimal performance. Addressing these challenges will be crucial for the widespread adoption and scalability of DEA-RNN in various real-world applications.

In conclusion, the DEA-RNN model represents a significant advancement in the field of cyberbullying detection on social media platforms like Twitter. Its high accuracy, robustness against diverse linguistic patterns, and superior performance metrics highlight its potential as an effective tool for online harassment mitigation. The model's practical applications and implications underscore its relevance, while future research directions provide a pathway for further enhancements. Despite some limitations, DEA-RNN sets a new benchmark in leveraging deep learning for social media monitoring, contributing to a safer and more inclusive online environment.

CONCLUSION

This paper developed an efficient tweet classification model to enhance the effectiveness of topic models for the detection of cyber-bullying events. DEA RNN was developed by combining both the DEA optimization and the Elman type RNN for efficient parameter tuning. Furthermore, it was tested in comparison with the existing Bi-LSTM, RNN, SVM, RF, and MNB methods on a newly created Twitter dataset, which was extracted using CB keywords. The experimental analysis showed that the DEA-RNN had achieved optimal results compared to the other existing methods in all the scenarios with various metrics such as accuracy, recall, F-measure, precision, and specificity. This signifies the impact of DEA on the performance of RNN. Although the hybrid proposed model obtained higher performance rates than the other considered existing models, the feature compatibility of DEA-RNN reduces when the input data is increased greater than the initial input. The current study was limited only to the Twitter dataset exclusively; other Social Media Platforms (SMP) such as Instagram, Flickr, YouTube, Face book, etc., should be investigated in order to detect the trend of cyber bullying. Then, the possibility of utilizing multiple source data for cyber-bullying detection will be investigated in the future. Furthermore, we performed. the analysis only on the content of tweets; we could not perform the analysis in relation to the users' behavior. This will be in future works. The proposed model works to detect cyber bullying utilizing textual content of tweets, whereas theother type of media such as images, video, and audio is still an open research area and future research directions. Besides, we aim to classify and detect CB tweets in a real-time stream.

REFERENCES

1. Agrawal, S., Awekar, A. (2018). Deep learning for detecting cyberbullying across multiple social media platforms. *arXiv preprint arXiv:1801.06482.

2. Zhang, Z., Luo, L., Yue, K. (2018). Detecting Offensive Language on Social Media to Protect Adolescent Online Safety. *Journal of Internet Services and Information Security (JISIS)*, 8(3), 60-72.

International Journal of Mechanical Engineering Research and Technology

ISSN 2454 – 535X

http://www.ijmert.com

Vol.16 Issue 2, 2024

3. Waseem, Z., Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. *Proceedings of the NAACL Student Research Workshop*, 88-93.

4. Mishra, P., Yannakoudakis, H., Black, A. W., Shutova, E. (2019). Tackling Online Abuse: A Survey of Automated Abuse Detection Methods. *Journal of Computational Linguistics*, 45(3), 455-495.

5. Davidson, T., Warmsley, D., Macy, M., Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)*.

6. Badjatiya, P., Gupta, S., Gupta, M., Varma, V. (2017). Deep Learning for Hate Speech Detection in Tweets. *Proceedings of the 26th International Conference on World Wide Web Companion*, 759-760.

7. Gamba, M., Romano, A., Bonchi, F. (2017). Community-based Approaches to Identification of Key Spreaders in Online Social Networks. *IEEE Transactions on Network and Service Management*, 14(2), 241-253.

8. Chen, Y., Zhou, Y., Zhu, S., Xu, H. (2012). Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. *Proceedings of the 2012 International Conference on Privacy, Security, Risk and Trust*, 71-80.

9. Risch, J., Krestel, R. (2020). Bagging BERT Models for Robust Aggression Identification. *Proceedings of the Fourth Workshop on Online Abuse and Harms (ACL 2020)*, 55-61.

10. Saha, K., Chandrasekharan, E., De Choudhury, M. (2019). Prevalence and Psychological Effects of Hateful Speech in Online College Communities. *Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2019)*.

11. Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., Vakali, A. (2017). Mean Birds: Detecting Aggression and Bullying on Twitter. *Proceedings of the 2017 ACM on Web Science Conference*, 13-22.

12. Fortuna, P., Nunes, S. (2018). A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys (CSUR)*, 51(4), 1-30.

13. Mozafari, M., Farahbakhsh, R., Crespi, N. (2020). A BERT-based Transfer Learning Approach for Hate Speech Detection in Online Social Media. *Proceedings of the Fourth Workshop on Online Abuse and Harms (ACL 2020)*, 1-10.

14. Rosa, H., Pereira, N., Ribeiro, R., Ferreira, P., Carvalho, J. P., Oliveira, M., Coheur, L., Paulino, P., Ribeiro, R., Trigo, P. (2019). Automatic Cyberbullying Detection: A Systematic Review. *Computers in Human Behavior*, 93, 333-345.

International Journal of Mechanical Engineering Research and Technology

ISSN 2454 – 535X

http://www.ijmert.com

Vol.16 Issue 2, 2024

15. Gamba, M., Romano, A., Bonchi, F. (2017). Learning and Analyzing Behavior in Online Social Networks. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1105-1114.

16. Zhang, Z., Luo, L., Yue, K. (2018). Detecting Offensive Language on Social Media to Protect Adolescent Online Safety. *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 334-337.

17. Cheng, J., Danescu-Niculescu-Mizil, C., Leskovec, J. (2015). Antisocial behavior in online discussion communities. *Proceedings of the Ninth International AAAI Conference on Web and Social Media (ICWSM 2015)*, 61-70.

18. Wulczyn, E., Thain, N., Dixon, L. (2017). Ex Machina: Personal Attacks Seen at Scale. *Proceedings of the 26th International Conference on World Wide Web (WWW 2017)*, 1391-1399.

19. Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., Chang, Y. (2016). Abusive Language Detection in Online User Content. *Proceedings of the 25th International Conference on World Wide Web (WWW 2016)*, 145-153.

20. Pavlopoulos, J., Malakasiotis, P., Androutsopoulos, I. (2017). Deep Learning for User Comment Moderation. *Proceedings of the First Workshop on Abusive Language Online (ACL 2017)*, 25-35.