International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 05 | May - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Deep Learning Algorithms for Cyber-Bulling Detection in Social Media Platforms

Tharun Koduri [1], T Narsimha Sai [2], P Shiva Sai Laxman[3], P Mohan[4], Mr B Mariya Joseph[5]

[1,2,3,4] *UG Scholars,* [5]*Assistant Professor*
[1,2,3,4,5] *Department of CSE[Artificial Intelligence & Machine Learning],*
[1,2,3,4,5] *Guru Nanak Institutions Technical Campus, Hyderabad, Telangana, India*

Abstract - *Information and Communication Technologies have driven social networking and communication, but cyber bullying is one major challenge. Manual and ineffective user-dependent cyber bullying reporting and blocking mechanisms exist. Conventional Machine Learning and Transfer Learning were examined for automatically detecting cyber bullying. The research made use of an extensive dataset and systematic annotation procedure. Textual, sentiment and emotion, static and contextual word embeddings, psycholinguistics, term lists, and toxic features were all utilized in the Conventional Machine Learning approach. The use of toxicity features for cyber bullying detection was introduced in the present research. Contextualized word Convolutional Neural Network (Word CNN) word embeddings exhibited similar performance, with the choice of when to use word embeddings for its higher F-measure. The textual features, the embeddings, and the toxicity features established new standards when individually input into the model. The model attained an improved F-measure with the combination of textual, sentiment, the embeddings, the psycholinguistics, and the toxicity features in the Logistic Regression model. This outcompeted Linear SVC in training time and processing high-dimensionality features. Transfer Learning made use of the base model Word CNN for the base-model fine-tuning, attaining the benefit of faster training calculation compared to the base-models. The cyber bullying detection through Flask web was also done, and an accuracy was achieved. Mention of the actual name of the dataset was avoided to respect its privacy.*

***Key Words*:** *Cyberbullying, deep learning, LSTM, social networks.*

1 INTRODUCTION

"With the emergence of online social networks, the easy access to information and communication technology, and the common use of computers and smartphones, Internet users are subject to increased levels of freedom of expression. Moreover, social media users often have the right to hide their identities, and in that case, the potential exploitation of different functionalities is made possible. The matter of offensive content has become very notable within the scope of social networking. Offensive content is anything that indicates abusive behavior with the aim of harming others. Different types of abusive content can be found on social networking websites, such as sexism, racism, cyber abuse, hate speeches, and toxic comments [1]. Hate speeches are becoming very common in face-to-face communication as well as online communication during the last few years [2]. Different aspects contribute to the occurrence of hate speeches. First of all, the anonymity that is provided online, particularly on social networks, tends to induce individuals to act in an aggressive manner [3]. Moreover, the new trend of speaking out online about one's views has also led to the propagation of hate speeches. As the spread of such discriminatory communication has several undesirable effects on society, governments and social networking websites can rationally benefit from the use of tools for the detection and prevention of hate speeches [4]. Deep Learning (DL) is an approach adopted within the area of machine learning to facilitate the execution of unsupervised learning by using un labeled data. Within the realms of data mining and the categorization of data, different studies in the field of investigation found the application of DL techniques to predict and categorize events such as hate speech detection and opinion categorization. Some of the types of Deep Learning Networks are Feed Forward Neural Networks, Deep Belief Networks, Convolutional Neural Networks (CNN), Restricted Boltzmann Machines, Recurrent Neural Networks (RNN), and Stacked Denoising Autoencoders [5]. Various methodologies have been examined in the effort to identify hate speech, including traditional classifiers, neural network-based classifiers, or the combination of both methodologies. The advancement in methodologies for word embeddings with deep learning has introduced tools such as Fast Text, Glove, word2vec, and transformer-based methodologies, which were used to obtain the retrieval of advanced representations. The pre-trained representation using representation learning tools provides resultant embedding-based representations, which can be implemented with traditional as well as advanced classifiers. This augments the range of methodologies that can be used to detect hate speech, offering an extensive variety of possible solutions to various real-world applications [6]. The rise in cyberbullying on social networking websites and the variety in its type has led to adverse impacts on the victim.

The adverse impacts that manifested on the victims upon facing cyberbullying are numerous, including adverse impacts on physical and mental health such as anxiety, depression, thinking, and low self-esteem, and on occasions, it led to suicide [27]. With the occurrence of adverse impacts and the rise in bullying on social media websites, there has been the need to obtain an approach to decrease and prevent the phenomenon of cyberbullying [28]. In a comparative work published by [29] on cyberbullying detection on social media for the period of the previous five years, introduced a set of earlier studies that employed machine learning and deep learning models in good endeavors to detect and categorize the phenomenon of cyberbullying. They concluded on the basis of what was observed in the outcomes of the related works to achieve improved performance in future studies. It is suggested to employ deep learning models (BiLSTM classifier and BERT) and in the case of using machine learning models prefer using SVM and NB as classifiers. From the survey of the earlier studies several limitations have been noted such as multi-class cyberbullying categories, experiments on large datasets for aggression detection, datasets from different multimedia platforms not considered. So the major objective of the present search was to construct the detection model that improves the performance of the classifiers on large common datasets through feature extraction techniques combination.

2 LITERATURE SURVEY

In [8], the model is suggested to signify the double meaning of cyberbullying using an innovative CNN concept for content analysis and an unethical method to manage by providing the organization with less accuracy. Compared to the other studies, the data collected are found to provide higher accuracy and categorization.

A systematic review of n=186 websites on the internet databases was published by [9]. In the paper, there were 10 literature reviews that were screened to assess and discuss data on the effectiveness of ML in preventing cyberbullying. The greater majority of the models to predict cyberbullying capitalize on content-based features. Although the most common are the support vector machines, naive Bayes, and convolutional networked networks, most of these are the features that are gleaned from social media postings. ML is a sophisticated preventive strategy which may enhance and combine adolescent education programs and act as the basis of the creation of technology-based automated screening procedures.

From [10], there is an approach to detect cyberbullying developed using fuzzy logic in which the interaction of the two users is continually watched and the emotive content of each message is ascertained. Depending upon the amount of emotional data that is involved, the behaviour is labeled as good or as bullying. The user's account is automatically canceled and reported whenever the amount of bullying observed exceeds the predetermined level. They concluded that when used together with social networking websites, it can be an efficient tool for preventing online harassment. The formulated algorithm can also be used for surveillance and monitoring human behaviour.

The new pre-trained BERT model was built by [11] and was assessed using two social media databases. One of the databases had a comparatively small network layer on top acting as the classifier, and the other database contained a bigger network layer on top acting as the classifier.

Deb and Aiyar [12] proposed the new model named DEA-RNN, an Elman-type recurrent neural network (RNNs) with enhanced dolphin allocation algorithm (DEA). The experimental results were such that DEA-RNN outperformed the others in different conditions with an overall accuracy of 90.45%, precision of 89.52%, recall of 88.98%, F1-score of 89.25%, and specificity of 90.94%.

[13] conducted a study to develop a model with the ability to identify cyberbullies in Bangla and Romanized Bangla messages through various machine learning and deep learning methodologies. On the other hand, the machine learning model multinomial naive Bayes surpassed others with an accuracy of 84% on the Romanized Bangla dataset and 80% on the overall dataset.

The authors of [14] proved that incorporating the TFIDF vectorizer on top of Farasa NLTK improved SVM's performance in cyberbullying classification compared to the NB classifier. The findings indicated that SVM still performed better in cyberbullying content detection compared to the regularized NB at an accuracy rate of 95.742%.

Study [15] used a sentiment detection system that utilized Recurrent Neural Networks (RNN) for the evaluation of the text and Convolutional Neural Networks (CNN) for the evaluation of images. The data from the text was obtained using the Twitter API. The functionality and efficacy of the models were demonstrated to achieve an accuracy between 0.951 and 0.911. Furthermore, the scores were found to be 0.910 and 0.890 for RNN and CNN, respectively.

The CNN and LSTM combination model using the Kears Functional API has been proposed by [17].After training the model, the accuracy rate of the image-based prediction is 86%, and the accuracy rate of the textual prediction is 85%.

3 PROBLEM STATEMENTS

In the age of mass digital communication, social media has been at the forefront of the ways in which people communicate, share views, and coalesce into communities. Yet, there is a darker side to such convenience -

cyberbullying has come to the fore. Cyberbullying is the abuse of digital media to harass, intimidate, or belittle people and is usually carried out anonymously. It has deep psychological and emotional implications, specially for teenagers and younger users. Inspite of the steep rise in such acts, current methods to detect and counter cyberbullying are mostly ineffective, relying heavily on user reports or simple filtering methods.

There have been attempts to apply conventional machine learning methods to automate the identification of cyberbullying but they have serious drawbacks. Conventional machine learning methods usually rely on manual feature extraction and cannot catch the emotional and context-specific nuances of language in social media discourse. As such, they tend to miss subtler implicit abuse—such as sarcasm, coded slurs, or passive-aggressive statements—due to high false negatives. In addition, they are not sufficiently adaptive to deal with shifting language patterns in social media.

One of the most significant concerns is the compute efficiency of existing deep learning models, particularly in working with large data sets. Training sophisticated models from scratch requires significant computational capacities, which add to the environmental expense (e.g., CO_2 footprint) and hinder scalability. In addition, the models are not fine-tuned to tradeoff performance with efficiency, which is critical in the context of real-time applications such as monitoring social media. The failure to leverage sophisticated linguistic attributes like toxicity scores, polarity of sentiment, emotional tone, and psycholinguistic indicators further compromises the efficiency of current tools for detecting cyberbullying. A recognition of explicit abuse is only possible with these attributes, which are necessary for identifying subtler and psychologically insidious language as well. Without the integration of such disparate inputs, the possibility to build resilient, generalizable models with the potential to adapt to various online media and user behaviours is foregone.

Hence, there is an immediate necessity for a smart, efficient, and effective solution to counter cyberbullying. This project is intended to solve all such issues through the implementation of deep learning (Word CNN) using transfer learning to enable it to both be efficient in its processing cost and accuracy. With the implementation of extensive text and linguistic features—such as toxicity and contextual embeddings—the system seeks to significantly improve its detection of social media cyberbullying content, rendering social media safer for its users.

4 PROPOSED METHODOLOGY

The model was built and developed as a sequential type using Keras and Python. Keras is an advanced neural network API which is compatible with the TensorFlow open-source machine learning framework [18]. This used the Natural Language Toolkit (NLTK), which is most widely popular and well-known as one of the leading NLP libraries in the Python universe. PyCharm was used for the whole implementation procedure. Through various experiments, the best conditions and outcomes there were established. The four-step proposed algorithm various steps of cyberbullying detection which are implemented the state.

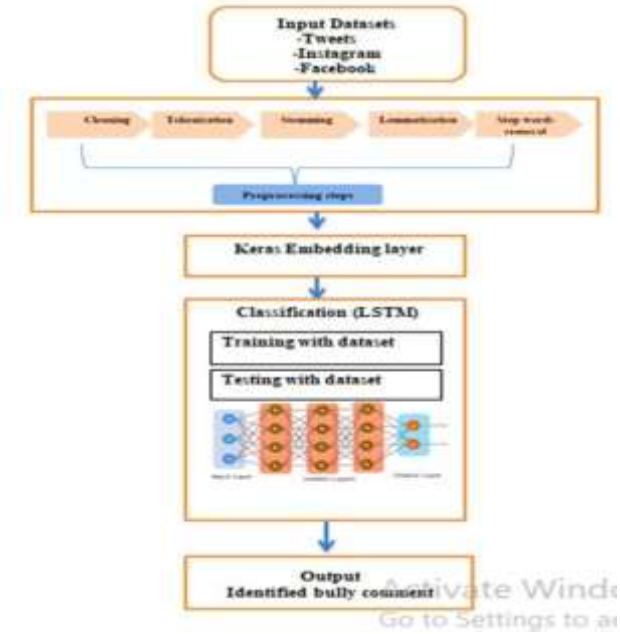

Fig1:- Architecture of The Proposed Models.

4.1 Datasets

In identifying cyberbullying textual content, the Hate-speech datasets were employed in the experiments of the present work, including the Twitter, Instagram, and Facebook datasets.

Twitter is a very popular site for microblogging and social networking, where users are able to share and present short messages labeled as "tweets." The dataset employed is found online at https://data.world/thomasrdavidson/hatespeech-and-offensive-language; the dataset consists of 24783 labeled as neither (1430), offensive language (19190), or hate (4163) tweets. The second dataset is the Instagram dataset, which is a known social network founded on media where users can upload, like, and comment on photos. It is found online at https://github.com/nurindahpratiwi/dataset-hate-speechinstagram/blob/master/572-hate-speech-dataset.csv and consists of 572 labeled as neither (231), offensive language (55) or hate (286) Instagram posts. The third employed dataset for the case of Facebook is found online at https://github.com/stjordanis/Dynamically-Generated-Hate-Speech-Dataset/blob/main/2020-12-31-Dynamically-Generated-Hate-Dataset-entries-v0.1.csv and consists of 40623 labeled as neither (5441), offensive language (16683), or hate (18499) Facebook comments. Each dataset consisted of a CSV file with two columns, where each row is made up of one comment or set of multiple consecutive

comments of one conversation and its related label, which for the employed hate dataset was [0, 1, 2], where 1 is labelled as hate speech, 2 as the case of offensive language, and 0 as the case of neither to train the classified consideration system.

4.2 Word Embedding

Once data preprocessing is done, model construction and training phase is initiated. The word embedding method is employed to create LSTM models. Word embeddings are various methods for creating numerical representations of text. Word Embedding possesses the remarkable characteristic of generating the same representations for words with similar semantic meanings. This characteristic helps a machine perceive the text to have meaning instead of merely viewing it as a series of random integers. With random initial weights, Keras's embedding layer was employed to train the embeddings as part of the network. This specific embedding choice (Keras) is remarkable because it is task-specific. By incorporating semantics as opposed to merely using features derived from raw text, word embedding improved the proposed model's accuracy level compared to other typical detection methods [20].

4.3 Algorithm

Long Short-Term Memory (LSTM) networks are specialized recurrent neural networks intended to learn and remember long-term relationships within sequences of data. As opposed to typical RNNs, LSTMs resolve the vanishing gradient problem, which prevents learning long-term relationships during backpropagation. This makes LSTMs very useful for sequence-related tasks like time-series prediction, natural language processing, and text classification. LSTMs are extensively adopted in areas where memory needs to be preserved within longsequences.

The fundamental behaviour of LSTM networks is dictated by gated architecture composed of input, output, and forget gates. They control information flow, deciding what to keep, forget, and output at every time step. Each gate has its operations expressed mathematically using inputs (xt), current and past hidden states (ht and ht-1), and learnable parameters (W and b). This design enables the network to selectively process and retain critical information in long sequences.

4.4 System Architecture

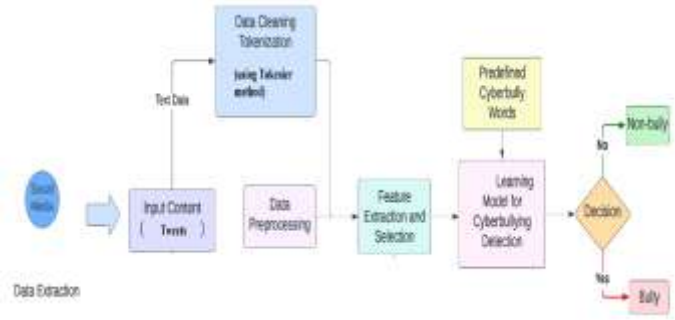

Fig2:System Architecture of the proposed system

4.5 Results

Cyberbullying detection has become a vital focus in the fight against online abuse, especially with the rise of social media. Unlike older methods that required manual feature selection, deep learning models like CNNs, RNNs, and LSTMs can understand the flow and meaning of conversations on their own. These models are especially good at handling huge amounts of unstructured text, making them more accurate and adaptable across different platforms. The use of hybrid models, like combining CNNs with LSTMs, allows for deeper insights by capturing both local and contextual patterns in language. Pre-trained word embeddings such as Word2Vec, GloVe, and fastText add more meaning to the models' understanding of text. Still, challenges like imbalanced data and transparency in model decisions remain. Even so, deep learning is paving the way for smarter, more human-centered systems that can help create safer, kinder digital spaces for everyone.

5. Future Enhancements & Conclusion

This study presents an efficient model for detecting cyberbullying, motivated by the limitations of previous approaches. By leveraging deep learning algorithms, the model eliminates the need for manual feature extraction, improving both accuracy and scalability. The use of three diverse datasets enhances the model's adaptability across different contexts, making it more effective in identifying bullying behaviour. A key improvement was replacing the sigmoid activation function with ReLU in the hidden layers, significantly boosting LSTM performance by allowing better learning from the training data. The study also emphasizes the importance of future enhancements, such as incorporating image and video analysis and utilizing multilingual datasets, to address cyberbullying more comprehensively across global platforms.

REFERENCES

[1]. P. Fortuna, and S. Nunes, "A survey on automatic detection of hate speech in text", ACM Computing Surveys (CSUR), Vol. 51, No. 4, pp. 1-30, 2018.
[2].FBI.2015.2015hatecrimestatistics.Retrievedfromhttps:// ucr.fbi.gov/hate-crime/.
[3]. P. Burnap, and M.L. Williams, "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, 2015.
[4]. M. Wendling, "The year that angry won the internet", BBC Trending, 2015.
[5]. B. Haidar M. Chamoun, A. Serhrouchni, "Arabic Cyberbullying Detection: Using Deep Learning", In the 7th International Conference on Computer and Communication Engineering (ICCCE). pp.284-289, 2018.
[6]. J.S. Malik, G. Pang, and A.V.D. Hengel, "Deep learning for hate speech detection: a comparative study", arXiv preprint arXiv, pp. 2202-09517, 2022.
[7]. M. H. Obaid, S. K. Guirguis and S. M. Elkaffas, "Cyberbullying Detection and Severity Determination Model," in IEEE Access, Vol. 11, pp. 97391-97399, 2023.
[8]. V. Balakrishnan, S. Khan, HR. Arabnia "Improving cyberbullying detection using Twitter users' psychological features and machine learning", Comput Secur, Vol. 90, pp. 101710, 2020.
[9]. G. Perasso, (2020) "Cyberbullying detection through machine learning: Can technology help to prevent internet bullying", Int J Manag Humanit, Vol. 4, pp. 57–69, 2020.
[10] S. Prashar, and S. Bhakar, "Real time cyberbullying detection", Int J Eng Adv Technol, Vol. 9, pp. 5197– 5201, 2019.
[11] J. Yadav, D. Kumar, and D. Chauhan, "Cyberbullying detection using pre-trained bert model", In 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), IEEE, pp. 1096-1100, July, 2020.
12] B.A.H., Murshed, J.Abawajy, S. Mallappa, M.A.N Saif., and H.D.E. Al-Ariki "DEA-RNN: A hybrid deep learning approach for cyberbullying detection in Twitter social media platform", IEEE Access, No. 10, pp.25857- 25871, 2022
[13] M.T. Ahmed, M. Rahman, S. Nur, A. Islam, and D. Das, "Deployment of machine learning and deep learning algorithms in detecting cyberbullying in bangla and romanized bangla text: A comparative study", In 2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), pp. 1-10, IEEE, 2021.
[14] A.M Alduailaj., and A. Belghith, "Detecting Arabic Cyberbullying Tweets Using Machine Learning", Machine Learning and Knowledge Extraction, Vol. 5, No. 1, pp. 29- 42, 2023.
[15] M. AGBAJE, and O. Afolabi, "Neural Network-Based Cyber-Bullying and Cyber-Aggression Detection Using Twitter Text", 2022.
[16] K. Shah, C. Phadtare., and K. Rajpara, "CyberBullying Detection in Hinglish Languages Using Machine Learning", International Journal of Engineering Research & Technology (IJERT), Vol. 11, Issue 05, May, 2022.
[17] V. Vijayakumar, P. Hari, and P.Adolf, "Multimodal Cyberbullying Detection using Hybrid Deep Learning Algorithms", International Journal of Applied Engineering Research, Vol. 16, No. 7, pp. 568-574, 2021.
[18] M.E Kula., "Cyberbullying: A Literature Review on Cross-Cultural Research in the Last Quarter", Handbook of Research on Digital Violence and Discrimination Studies, pp.610-630. 2022.
[19] E. Bashir, and M. Bouguessa, "Data Mining for Cyberbullying and Harassment Detection in Arabic Texts", International Journal of Information Technology and Computer Science, Vol. 13, No. 5, pp. 41-50, 2021.