Using Deep Learning LSTM and CNN with Word Embedding for the Detection of Offensive Text on Twitter

Abdulkarim Faraj ALQAHTANI
Department of Electrical Engineering and Computer Science
Florida Atlantic University, Boca Raton, FL, USA

Mohammad ILYAS
Department of Electrical Engineering and Computer Science
Florida Atlantic University, Boca Raton, FL, USA

ABSTRACT

Reliance on technology has become prevalent in many aspects of our life. Technology provides many benefits in people's lives, but there are also many concerns generated by this technology. Social media provides benefits, as it allows people to express their comments, opinions and feelings. However, social media can create an inappropriate environment that generates hate speech, offensive text and cyberbullying. Offensive text is included in comments or tweets that pass between users of social media; thus, this is a serious issue that needs to be determined and detected by employing natural language processing. This paper proposes an automatic cyberbullying method for the detection of offensive text using two deep learning models to provide high accuracy. The models employed in this paper are long short-term memory and convolutional neural network (CNN), which are used to classify whether tweets contain offensive text or non-offensive text. In addition, in this paper we compare the CNN model with prior papers that used the same model to show the improvements in accuracy we obtained. We combined five hate speech datasets that contained 162 k tweets to perform the detection in our models. The highest accuracy of our models was approximately 93%, indicating promising results. Our method was found to be more effective at detecting offensive text than the existing method when tested on combined datasets.

Keywords: Cyberbullying, Deep Learning Models, Natural Language Processing, Sentiment Analysis, Social Media, Word Embedding.

1. INTRODUCTION

Bullying has long existed, but because it happened physically there was no form of monitoring that could stop it. Nowadays, bullying takes place through social media (SM) platforms, which is known as cyberbullying. SM involves many people who are diverse in age, gender, religion and color. Users of SM can communicate with each other, and they may hurt others by writing negative comments or tweets that include bullying words. This behavior most likely involves users of an early age, especially teens, which makes this situation very concerning.

Cyberbullying may lead to risk issues for victims, so they may harm themselves when they read negative writing against them while sharing their opinions or photos on social networks. Cyberbullying can reach victims through various methods such as text, calling or sharing photos or videos. Studies have shown that cyberbullying may have serious impacts on victims, including mental health issues, anxiety, depression and suicide. Therefore, cyberbullying negatively impacts society, so it is important to discover solutions that can limit this behavior. The authors in [5] emphasize that due to the harmful effects of cyberbullying on its victims, it is crucial to find effective ways to detect and prevent it. Machine learning can be a useful tool for this purpose because it can analyze data and generate models that automatically classify appropriate actions. By studying the language patterns of cyberbullies, machine learning can identify instances of cyberbullying and inform the development of strategies to prevent it.

Cyberbullying is present in many environments such as communities, workplaces and schools. Studies have demonstrated the impact of cyberbullying in schools and emphasized that this behavior destroys friendships and negatively influences school activities and schoolwork. Cyberbullying can also cause physical and mental health issues as well as contribute to a lack of confidence of victims. According to a study conducted in the United States, most students aged 9 to 12 have had experiences of cyberbullying [1]. This study demonstrates how extensive this behavior is and that online communication is not safe for tweens. In addition, this study presents prior studies that demonstrate the dangers of cyberbullying and how it can cause victims to experience suffering and negatively impact their schoolwork. In the United States, it has been found that a significant number of teenagers (over 40%) have experienced cyberbullying. In this study, we suggest using deep learning models as a way to identify and address instances of cyberbullying [3]. With the increase of SM users in recent times, this number may represent several billion people across the world who use social network platforms for communication [2]. This behavior will only become more prevalent with the increasing use of SM, so a technical way to determine and decrease this issue is needed.

The detection of cyberbullying is considered a challenging task, and it may be impossible to be detected by human beings alone, especially when dealing with big data. Thus, by using NLP techniques it will be possible to detect offensive texts, though these techniques vary, and their use depends on the size of the data and the text detection technique employed. NLP techniques require feature selection, preprocessing steps and machine learning or deep learning algorithms to detect text and reduce the effects of cyberbullying. Thus, there are various NLP methods for detection, depending on the kind of data being analyzed. To detect textual data, sentiment analysis can be used for understanding human languages and processing the text. Sentiment analysis involves categorizing text as positive,

ISBN: 978-1-950492-70-1
ISSN: 2771-6368

https://doi.org/10.54808/ICSIT2023.01.86

negative, or neutral. Sentiment analysis can be used to monitor the emotions and opinions of the online community towards people, events, and other topics. This information can be used to gauge the public's mood or sentiment [4].

In this paper, we test two deep learning models named long short-term memory (LSTM) and convolutional neural network (CNN) to improve the accuracy of the detection of offensive text in Twitter's dataset. Our models provide higher accuracy than prior work, and we have increased the number of records of the dataset by combining five datasets to ensure the deep learning models provide high accuracy when the dataset is large. Thus, the main contribution of this paper is to propose a deep learning approach for detecting cyberbullying and improving accuracy compared with prior work. Also, we combine five datasets related to offensive text or non-offensive text comprised of around 162 thousand tweets to test the effectiveness of our detecting models.

Background of Deep Learning and Cyberbullying

Deep learning is one of the parts of machine learning. Deep learning uses algorithms to mimic the human brain, which are called artificial neural networks. In addition, deep learning algorithms are able to learn and understand the important features and patterns in data to improve performance. Therefore, deep learning algorithms have been proven to be able to classify and analyze all kinds of large data; however, more performance tools are needed for analysis such as powerful hardware and central processing units. This paper uses deep learning techniques for determining and mitigating this behavior in online platforms with using to achieve optimal accuracy. As cyberbullying is concern issue, the automatic system can help to detect cyberbullying, because it is important to know that the impacts of bullying can lead to straggly situations for the victims. Thus, this will lead and encourage to apply automatic system to find safe environment in the online platforms.

Paper Contributions

The contributions of this paper are summarized below:

- Combined five datasets are related to offensive and hate speech text in the Twitter platform.
- Use oversampling technique to make the dataset balanced.
- Apply the word embedding feature.
- Two deep learning models are applied for text classification.

2. LITERATURE REVIEW

This section reviews prior papers that discussed the detection of cyberbullying on social networks using various techniques with deep learning and machine learning algorithms. The authors in [5] used the multichannel technique of three different deep learning models (transformer block, BiGRU, and CNN) to make a final prediction. They combined three datasets from different sources, and the total records of the dataset were 55,788 tweets, which were labeled as either offensive or non-offensive. After that, they executed the preprocessing task, which included tokenization and vectorization, to convert the text to integers. For their experiment, they split the dataset multiple times, starting with 75% training and 25% testing and changing to 50% training and 50% testing and then 30% training and 70% testing. The reason for doing this is to improve the accuracy of their models. Thus, their proposed method had an accuracy rate of approximately 88%.

Different methods have been used in many papers to detect offensive text on social networks for the purpose of increasing the accuracy of detection. The authors in [6] suggested an approach that combines feature subset selection with deep learning for cyberbullying detection and categorization (FSSDL-CBDC). This suggested technique involved many phases, including preprocessing text, feature selection and classification. The models they used are the Salp Swarm Algorithm (SSA), which is used to describe and detect cyberbullying on social media platforms through a deep belief network (DBN). For feature subset selection, they used a binary coyote optimization-based feature subset selection (BCO-FSS), which helped increase the performance of classification. For the preprocessing phase in their experiment work, they employed a lexical normalization technique for deep cleaning in the dataset. Also, they used a spell corrector tool to delete unwanted vocabulary words, punctuation marks and missing values in the input data and also to improve the correction of spelling. For experiment results, the SSA-DBN model had the highest accuracy compared with the other algorithms, with a 99.983% accuracy rate. Figure 1 shows the processes for their experiment.

In addition, for various mothers to detect cyberbullying, the authors in [7] developed a model by combining Elman-type recurrent neural networks (RNNs) with the Dolphin Echolocation Algorithm (DEA) to get the optimal classification for the detection of cyberbullying on the Twitter platform. They emphasized that their developed model achieved superior results in all scenarios for their experiment, where it achieved an average of 90.45% as the highest result. In addition, they used their own new dataset collected from Twitter based on keywords that indicate cyberbullying. The dataset was comprised of 10,000 tweets that were labeled manually either as ‘‘0’’ for non-cyber bullying or ‘‘1’’ for cyberbullying. For preprocessing and data cleaning, in this step, various types of noise were removed, including URLs, hashtags and mentions, punctuation and symbols, and emoticons were transformed. Also, Tweets were transformed by converting them to lower case and stemming, tokenizing, and filtering stop words including spell checking. For feature extraction, Word2Vec and TF-IDF were used to extract features, with nouns, pronouns, and adjectives being the primary features, and adverbs and verbs providing additional information. They evaluated the performance of their developed model by comparing it with other algorithms such as Bi-directional Long Short-term Memory, RNN, Support Vector Machine (SVM), Multinomial Naive Bayes, and Random Forests that were utilized through their dataset. They confirmed that the DEA-RNN model achieved the highest accuracy.

As cyberbullying is related with sentiment analysis, the authors in [8] tested an appropriate approach by using deep learning techniques to detect sarcastic tweets on the Twitter platform. They used two models to classify their dataset: RNN and LSTM. They also used word embedding as feature selection to achieve high accuracy to provide an easy solution for companies that need to analyze reviews, comments and tweets of customers. For their system approach, they split the dataset into training and testing datasets to train the model, and the training dataset was executing by preprocessing steps. They implemented preprocessing steps, which included cleaning text and removing unwanted words, punctuation and stop words. Also, they applied word embedding and converted words to vectors, to feed LSTM model as a vector matrix. Finally, the classification step was done, and the highest result achieved in this approach was 88% accuracy with 15 epochs.

3. METHODOLOGY

Dataset: the dataset used in this system was taken from five different sources [9-13], which are these datasets collected from the Twitter platform and contain tweets related to offensive text, cyberbullying and hate speech. These datasets have been combined as one dataset comprised of 160,000 tweets that are categorized as either offensive tweets or non-offensive tweets. We combined these datasets into one dataset because there is a lack of datasets that involve large records for cyberbullying appropriate for the evolution of the model [5].

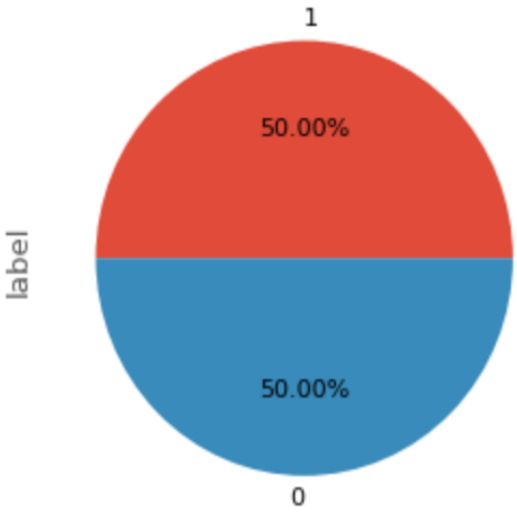


Figure 1: Show the dataset is balanced.

Proposed Approach: the proposed method for identifying the cyberbullying in tweets involves using both CNN and LSTM models. Thus, this system for the detection of cyberbullying has four main parts: 1) dataset 2) preprocessing 3) word embedding feature and 4) run the models. The dataset has been split into two datasets: the training dataset and testing dataset after they were preprocessed. The pre-processing step involved removing punctuation marks, special characters, links, and any other elements that do not contribute to determining whether a tweet is offensive. The word embedding feature has been applied in the pre-processing step, which can help identify the meaning of tweets. Thus, the training dataset has been prepared to be a vector matrix, which is used by the models to make it easy to detect offensive text. The final step is classification, which is applied by two models, and the dataset is tested to identify the accuracy of the detection. Figure 2 describes the overview of the workflow of our proposed approach.

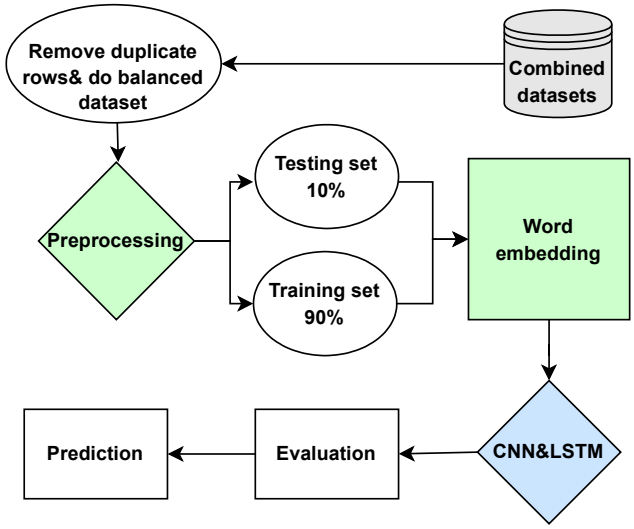


Figure 2: The proposed approach.

Preprocessing: the dataset is labeled as either 0 or 1, where 0 indicates positive tweets and 1 indicates tweets containing offensive words. Thus, the first step checks if there are any duplication records, deletes them, and checks the balance of the dataset. For our experiment, the dataset was imbalanced, and we used a function that balanced the dataset to make it have the same number of records, as Figure 1 shows. For the technique of balanced that we used is named oversampling which use to choose respondents in such a way that certain groups are overrepresented in the survey sample compared to their proportion in the overall rows. The reason of using this technique because the size of rows that belong to class 1 in the dataset was less than the rows that belong to class 0 as much around 40%. If the dataset remain imbalance, it will impact on the performance of model, and will classify based on the majority class only which will achieve low accuracy because the precision and recall achieve low results which impact on the overall accuracy. After the duplication and balance checks, the deep cleaning of tweets was executed to create the model. Thus, the clean text includes removing stop words, punctuation marks, spaces, specific characters, noisy and redirecting links as well as applying tokenization and stemming features. After the text was cleaned, the dataset was split into training and testing data. The word embedding feature was applied so the tweets were transformed into a sequence and then utilized for a word embedding model. Word embedding is used to present words in a numerical format, called word embedding, Figure 3 shows after the preprocessing is executed, and the text became cleaned, word embedding features are applied before run the models.

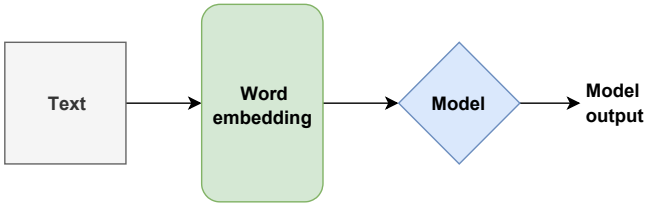


Figure 3: Word embedding.

All these steps are required for feeding the model to achieve the appropriate accuracy. In the last step, the data are ready to process, so it is then time to run the model to detect offensive

tweets. The tweets are transformed into word vectors and then input into a CNN, which will learn patterns in the data through repeated iterations. Also, we run LSTM to evaluate the result of the CNN.

Text Classification: in the dataset, the tweets have texts that need to be trained, so text classification is able to assign and organize the text that includes words that are determined to belong to the assigned class. This means that the models applied in our experiment are trained from the rows of the dataset and the labels of each row are organized. Thus, the rows that contain some words related to offensive words are labeled as 1, which means the tweet contains offensive text; otherwise, it is labeled as 0, which means the tweet contains positive text.

Models: 1- Long Short-Term Memory (LSTM) Networks: LSTM is a particular type of RNN model that has the ability to learn long-term dependencies. Information that is acquired over long periods of time is challenging to learn, so LSTM is designed to solve this issue. As a brief explanation of the mechanism of LSTM, the information will be controlled from one cell to another cell, so there is input and output with an internal cell that performs some processes to analyze and classify the input term as show in Figure 4.

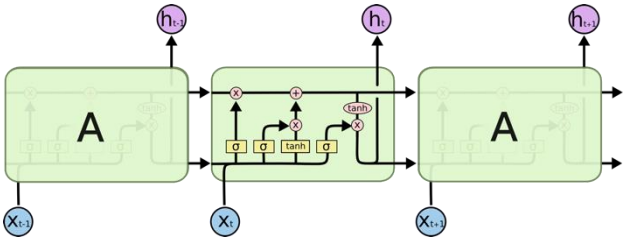
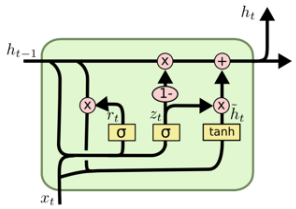


Figure 4: The repeating module in an LSTM contains four interacting layers [14].

Figure 5 shows the input cell, the internal cell, and the output cell. In the internal cell, Xt is the current input and the relation between previous cells, which saves on memory for previous input. Also, the output cell is also saved for the next coming input cell. Thus, the internal cell, which is called the forget gate, which is a sigmoid function, regulates the flow of information through the cell state, resulting in the discarding of certain information from the previous state. The figure below describes the satisfied equation for the forget cell in LSTM.



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

Figure 5: Gated Recurrent Unit [14].

2- Convolutional Neural Network (CNN): CNN is one of the optimal methods to use to extract the features of data in deep learning. CNN uses interconnected neurons that receive inputs, processes them by considering the weighted sum and applies an activation function, and then outputs the result to the next neuron. Each layer in a neural network, attempts to identify patterns or useful information within the input data as illustrated in Figure 6.

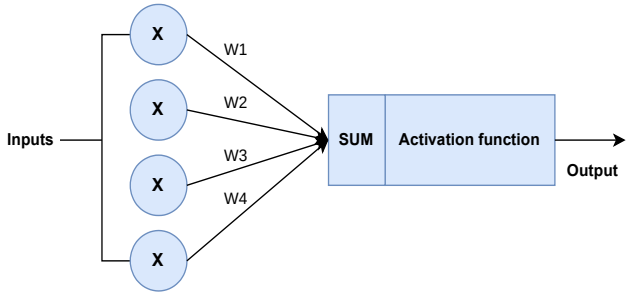


Figure 6: Convolutional Neural Network (CNN).

4. RESULTS

For the result achieved in our experiment, the accuracy achieved is similar for both LSTM and CNN. However, the time for running the LSTM model was faster than for CNN. Each model has the same quantity of spiting dataset, with 90% training and 10% testing. Overall, the accuracy was higher than some work prior using the same three datasets that were using in their work. There are some equations that use to calculate the accuracy, precision, recall and F1 scour. Thus, the equation that use to get the result by

$$\text{Accuracy} = \frac{TP + TN}{TN+TP+FN+FP}$$

Where: TP = true positive.
TN = true negative.
FP = false positive.
FN= false negative.

Which means in our experiment the total the number of offensive tweets that have been classified to class's offensive and total the number of non-offensive tweets that have been classified to class's non-offensive dividing on all total number of predictions.

Precision = (TP)/(TP+FP)
It is calculated as the number of true positive predictions divided by the total number of positive predictions made (TP + FP). The precision measures the accuracy in our experiment by correctly identifying offensive tweets in a test set, out of all tweets classified as offensive, both correctly and incorrectly.

Recall= (TP)/(TP+FN)
Recall measures the percentage of the true offensive tweets that have been classified correctly in a test set, among of all offensive tweets in the test set.

F1 Score = 2*((precision*recall) / (precision+recall))
F1 score is determined by dividing the product of precision and recall by the sum of precision and recall.

The effectiveness of our approach has been achieved to be 92.59% by LSTM, superior to the second accurate algorithm which is 91.64% that has been achieved by CNN. Our method was also analyzed using four other evaluation measures, precision, recall, F1-score and a confusion matrix. Our method demonstrated strong performance with 92% precision, 93% recall, and 93% F-score in identifying non-offensive and 93% precision, 92% recall, and 92% F-score in identifying offensive, as illustrated in Table 1.

Table 1: Summary of results using two models.

Model		Precision	Recall	F1 Score	Accuracy
LSTM	**0**	0.92	0.93	0.93	0.9259%
	1	0.93	0.92	0.92	
CNN	**0**	0.91	0.93	0.92	0.9164%
	1	0.93	0.90	0.92	

Moreover, in our experiment we have compared our results with other prior experiments that used various frameworks or methodologies that detect offensive and cyberbullying texts in the online platforms. As table 2 illustrates our experiment have been achieved higher effectiveness score.

Table 2: Our results compared to recent experiments of approaches that are used to detect offensive tweets and cyberbullying on Twitter.

Citation	Year	Models	Word Embedding	Accuracy
[5]	2021	Transformer block		87.99%
		CNN	Not Reported	87.28%
		BiGRU		87.43%
[7]	2022	Bi-LSTM		89.47%
		RNN	Not Reported	88.95%
		SVM		87.26%
		RF		88.33%
		MNB		86.83%
		DEA-RNN		90.94%
Ours	2023	LSTM	Applied	92.59%
		CNN		91.64%

Confusion Matrix: A confusion matrix is a tool used to evaluate the performance of a classification model by summarizing the number of correct and incorrect predictions. It provides a breakdown of the number of predictions for each class, which is useful for understanding the model's performance. It is a key element in evaluating classification models. Thus, the confusion matrix based on our experiment which we used to evaluate our models shown in Figure 7.

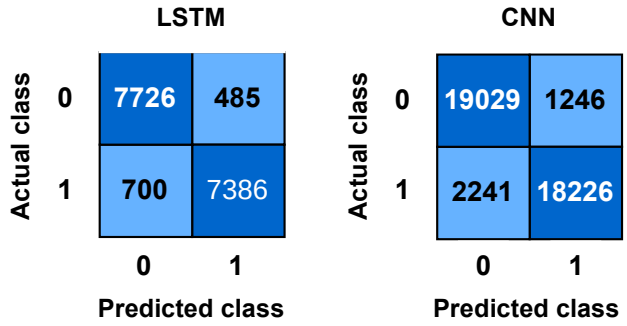


Figure 7: Confusion matrix for the performance of our classification.

5. CONCLUSIONS

To classify tweets as "offensive" or "non-offensive," a classifier, also known as a classification framework, analyzes a tweet for properties like patterns and words. In this paper, we combined five different datasets that contain tweets related to offensive text and hate speech from the Twitter platform. We also reviewed some prior works related to detecting cyberbullying text through various methods applied in these papers to show the improvement of performance. In our experiment, we applied two deep learning models: LSTM and CNN to classify the offensive text in the dataset we used. Although LSTM is more commonly used for detecting text and achieved higher accuracy, CNN achieves valuable accuracy. Also, we chose CNN, as the prior papers we targeted to compare with used CNN. Our experiment achieved higher accuracy compared with prior work because of certain features we used. Our dataset was imbalanced, so we used oversampling technique to balance it to help the models perform prediction and training easily, which helped improve the accuracy. We combined five datasets to increase the number of records in the dataset as deep learning models need big data when analyzing. Finally, we applied the word embedding feature, which converted tweets so they could be transformed into a sequence to be utilized for the models.

6. REFERENCES

[1] Patchin, J. W., & Hinduja, S. (2022). Cyberbullying among tweens in the United States: prevalence, impact, and helping behaviors. The Journal of Early Adolescence, 42(3), 414-430.

[2] Slonje, R., & Smith, P. K. (2008). Cyberbullying: Another main type of bullying?. *Scandinavian journal of psychology*, *49*(2), 147-154.

[3] S. Balakrishna and M. Thirumaran, "Programming Paradigms for IoT Applications: An Exploratory Study", In: Solanki, V. (Ed.), Díaz, V. (Ed.), Davim, J. (Ed.) Handbook of IoT and Big Data. Boca Raton: CRC Press, Taylor & Francis Group, Print. February 2019. doi: https://dx.doi.org/10.1201/9780429053290 2.

[4] A. Ikram, M. Kumar and G. Munjal, "Twitter Sentiment Analysis using Machine Learning," *2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 2022, pp. 629-634, doi: 10.1109/Confluence52989.2022.9734154.

[5] Alotaibi, M., Alotaibi, B., & Razaque, A. (2021). A multichannel deep learning framework for cyberbullying detection on social media. *Electronics*, *10*(21), 2664.

[6] Chandrasekaran, S., Singh Pundir, A. K., & Lingaiah, T. B. (2022). Deep learning approaches for cyberbullying detection and classification on social media. *Computational Intelligence and Neuroscience*, *2022*.

[7] 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," in IEEE Access, vol. 10, pp. 25857-25871, 2022, doi: 10.1109/ACCESS.2022.3153675.

[8] Salim, S. S., Ghanshyam, A. N., Ashok, D. M., Mazahir, D. B., & Thakare, B. S. (2020, June). Deep LSTM-RNN with word embedding for sarcasm detection on Twitter. In *2020 international conference for emerging technology (INCET)* (pp. 1-4). IEEE.

[9] DataTurks. Kaggle. Tweets Dataset for Detectionhttps://www.kaggle.com/dataturks/dataset-for-detection-of-cybertrolls (accessed on 15 January 2021).

[10] Davidson, T.; Warmsley, D.; Macy, M.; Weber, I. Automated hate speech detection and the problem of offensive language. *arXiv* 2017, arXiv:1703.04009.

[11] Elsafoury, F. (2020). Cyberbullying datasets. Mendeley. Available online https://www.kaggle.com/datasets/ saurabhs hahane/cyberbullying-dataset,[Accessed:04-Summer-2021].

[12] Wajid Hassan Moosa, & Najiba. (2022). Multi-lingual HateSpeech Dataset [Data set]. Kaggle. htt ps ://doi.org/10.34740/KAGGLE/DS/ 2260058.

[13] Zaidi, S. A. R. (2021, February 17). Suspicious tweets. Kaggle. Retrieved January 16, 2023, from https://www.kaggle.com/datasets/syedabbasraza/ suspicious-tweets.

[14] Christopher, O. (2015, August 27). Understanding LSTM networks. Understanding LSTM Networks -- colah's blog. Retrieved February 28, 2023, from https://colah.github.io /posts/2015-08-Understanding-LSTMs/