International Journal for Modern Trends in Science and Technology, 9(05): 353-358, 2023

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST0905058
Available online at: http://www.ijmtst.com/vol9issue05.html

Detection of Cyberbullying in Social Media using Machine Learning

Gowri J[1] | Vaishali V G[2] | Dinesh J[2] | Dharani Priya S[2] | Sanjay Madhan S[2]

[1]AssistantProfessor,Msc. SS, Sri Krishna Arts and Science College
[2]Msc. SS, Sri Krishna Arts and Science College

To Cite this Article
Gowri J, Vaishali V G, Dinesh J, Dharani Priya S and Sanjay Madhan S. Detection of Cyberbullying in Social Media using Machine Learning. International Journal for Modern Trends in Science and Technology 2023, 9(05), pp. 353-358. https://doi.org/10.46501/IJMTST0905058

Article Info
Received: 06 April 2023; Accepted: 11 May April 2023; Published: 15 May 2023.

ABSTRACT

Cyberbullying is a significant issue on the internet that affects both adults and teenagers. Mistakes like despair and suicide have resulted from it. A increasing demand exists for the regulation of material on social media platforms. The work that follows builds a model based on the detection of cyberbullying in text data using natural language processing and machine learning utilising data from two different types of cyberbullying, hate speech tweets from Twitter and comments based on personal assaults from Wikipedia forums. To determine the most effective method, three feature extraction techniques and four classifiers are examined. The model offers accuracy levels above 90% for data from Tweets and accuracy levels above 80% for data from Wikipedia[1].Tweets are examined by the process of NLP in machine learning and using some algorithms of deep learning.

KEYWORDS:Cyberbullying, NLP, detection,hatespeech,twitter.

1. INTRODUCTION

Online information sharing is popular among the millions of young people who spend their time on social networking sites. Social networks enable communication and information sharing with anybody, at any time, and with a large group of individuals all at once. Globally, there are more than 3 billion users of social media[2]. Bullying has been prevalent since the beginning of time, It's just the ways of bullying which have changed over the years , from physical bullying to cyberbullying, There hasn't been an effective measures to curb social bullying and it has become one of the alarming issues in recent times .

The National Crime Security Council (NCPC) defines cyberbullying as the deliberate harm or public humiliation of another person while using a mobile device, a video game app, or any other method to communicate or send text, photographs, or videos online. Cyberbullying can occur at any time, any day of the week, and you can contact anyone online. Cyberbullying can take the form of text, images, or video that is posted in an anonymous way. Finding the author of this post is sometimes impossible and can be complicated. Also, it was impossible to delete these messages in the future. Social media is primarily used for text or image-based cyberbullying. A system can respond appropriately if bullying text can be discriminated from non-bullying text. For social media platforms and other messaging services, an effective cyberbullying detection system can be helpful in

thwarting such attacks and lowering the incidence of cyberbullying. The cyberbullying detection system's goal is to locate the cyberbullying text and take into account its meaning. One first analyses the various aspects of a specific text before applying previous information or visuals to determine the context of the text. There is a need to develop a personalised system that can effectively and efficiently access such a text[2].

STRUCTURE OF PAPER

The paper is organized as follows: In Section 1, the introduction of the paper is provided along with the structure, important terms, objectives and overall description. In Section 2 we discuss related work. In Section 3 we have the complete information about Detection. Section 4 shares information about the proposed methodology and the results analysis.Section5 tells us about the future scope and concludes the paper with acknowledgement and references.

OBJECTIVES

. A system can respond appropriately if bullying text can be discriminated from non-bullying text. For social media platforms and other messaging services, an effective cyberbullying detection system can be helpful in thwarting such attacks and lowering the incidence of cyberbullying. The cyberbullying detection system's goal is to locate the cyberbullying text and take into account its meaning. One first analyses the various aspects of a specific text before applying previous information or visuals to determine the context of the text.

2. RELATED WORK

M. Di Capua et al. [3] propose an unsupervised approach to developing a cyberbullying model based on a combination of traditional textual features and some "social features." The features were classified as Syntactic features, Semantic features, Sentiment features, and Social features. As the input layer, the author used a Growing Hierarchical Self Organizing Map (GHSOM) network with a grid of 50 x 50 neurons and 20 features. M. Di Capua et al. used the k-means clustering algorithm, in conjunction with GHSOM, to classify the Form spring dataset. The outcomes of this hybrid unsupervised methodology outperformed the previous results. The author then applied three different Machine Learning Models to the YouTube dataset: a Naive Bayes Classifier, a Decision Tree Classifier (C4.5), and a Support Vector Machine (SVM) with a Linear Kernel. When comparing the Form Spring tests to the YouTube dataset, it was discovered that clustering results for hate posts had a lower precision, as textual analysis and syntactical features performed differently on both sides. This hybrid approach produced a low recall and F1 Score when applied to the Twitter dataset. The authors' model can be improved and used to create constructive applications to address cyberbullying issues.

J. Yadav et al.[4] propose a novel method for detecting cyberbullying in social media platforms by employing the BERT model with a single linear neural network layer on top as a classifier. The model is trained and tested using data from the Form spring forum and Wikipedia. The proposed model outperformed previous models with a performance accuracy of 98% for the Form spring dataset and 96% for the Wikipedia dataset. Because of the large size of the Wikipedia dataset, the proposed model produced better results without the need for oversampling, whereas the Form spring dataset required oversampling.

R. R. Dalvi et al.[5] propose using Supervised classification Machine Learning algorithms to detect and prevent Internet exploitation on Twitter.The live Twitter API is used in this study to collect tweets and create datasets. On the collected datasets, the proposed model employs both Support Vector Machine and Naive Bayes. They used the TFIDF vectorizer to extract the feature. The results show that the accuracy of the cyberbullying model based on the Support Vector Machine is nearly 71.25%, which is higher than the accuracy of the Naive Bayes, which was nearly 52.75%.

The goal of Trana R.E. et al. [6] was to create a machine learning model that minimised special events involving text extracted from image memes. The author has compiled a database of approximately 19,000 text views from YouTube. This study compares the effectiveness of the three machine learning machines used on the YouTube database, the Uninformed Bayes, the Support Vector Machine, and the convolutional neural network, with the results of the existing Form databases. Compares the results to the existing Form databases. The authors looked into algorithms for Internet cyberbullying in subcategories of the YouTube database. In the four categories of race, ethnicity, politics, and generalism, Naive Bayes outperformed SVM and CNN.

SVM outperformed the inexperienced Nave Bayes and CNN in the same gender group, and all three algorithms performed equally well with central body group accuracy. This study's findings provided data that can be used to differentiate between incidents of abuse and non-violence. Future work could concentrate on developing a two-part segregation scheme to test the text extracted from images to determine whether the YouTube database provides a better context for aggression-related clusters.

N. Tsapatsoulis et al. [7] present a comprehensive review of cyberbullying on Twitter. The significance of identifying various abusers on Twitter is emphasised. The paper thoroughly describes the various practical steps required for the development of an effective and efficient application for cyberbullying detection. The trends in data platform categorization and labelling, machine learning models and feature types, and case studies that used such tools are discussed. This paper will serve as the project's first step in detecting cyberbullying using machine learning.

Natural Language Processing (NLP) and Machine Learning were used to construct a cyberbullying detection model, according to G. A. León-Paredes et al. [8]. (ML). By combining Nave Bayes, Support Vector Machine, and Logistic Regression machine learning algorithms, a Spanish cyberbullying prevention system (SPC) was created. Twitter was mined for the dataset that was used in this study. Three strategies were employed to attain the highest accuracy of 93%. The accuracy of the incidents of cyberbullying identified by this technique ranged from 80% to 91% on average. NLP stemming and lemmatization techniques can be used to boost the system's accuracy even more. If feasible, a similar technique can be used to detect words in both English and regional languages.

P. K. Roy, et al. [9] provide specifics on how to develop a deep convolutional neural network-based request for the identification of hate speech on Twitter. On Twitter, hate speech-related tweets have been identified using machine learning algorithms like Logistic Regression (LR), Random Forest (RF), Naive Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), Gradient Boosting (GB), and K-nearby Neighbors (KNN), and their features have been eliminated using the tf-idf procedure. SVM was the best machine learning model, and in a 3:1 dataset used to test the train, it correctly predicted 53% of hate speech tweets. The uneven data was the cause of the poor prediction scale. The methodology is based on the forecasting of tweets that contain hate speech. Convolutional Neural Network (CNN), Long-Term Memory (LSTM), and their Contextual LSTM (CLSTM) combinations are used in advanced learning techniques to get the same results as a separate distributed database. The proposed DCNN model was used with 10-fold cross-validation, which led to a very high recall rate. For hate speech, it was 0.88, while for non-hate speech, it was 0.99. The k-fold cross-validation technique is a superior choice when there are unequal amounts of data, according to test results. The existing database can be expanded in the future for greater accuracy.

S. M. Kargutkar, et alsystem .'s to provide a double characterisation for cyberbullying was first proposed by them [10]. The system employs Convolutional Neural Network (CNN) and Keras for content analysis because the appropriate approaches at the time offered a less precise, guideless picture. Twitter and YouTube data were used in this study. Inaccuracy on CNN was 87%. Deep learning-based models have emerged that can detect instances of online abuse, overcome the limitations of conventional models, and increase adoption.

3. PROPOSED METHODOLOGY

To detect cyberbullying on social media it was not just done by sentimental analysis but also it’s proposed by the semantic analysis, syntactic analysis, sarcastic nature of the tweets in the platform, to perform the detection of bullying texts in tweets the first step begins with the traditional sentimental analysis , in that contextual mining is performed for identifying and extracting the texts with the help of subjective information to understand the motive, emotions and the opinion of tweets later the extracted texts comes under the testing and guiding process for the detection of cyberbullying on social media platform called Twitter.

The detection of cyberbullying in social media majorly processed by using Machine Learning and Deep Learning process ,In this for the increased performance of detection and extraction of texts Neural networking techniques were also implemented for the Accurate detection of

cyberbullying texts from the extracted datasets, The first step for the detection of the cyberbullying in the tweets is to extract datasets from the source like the list of tweets with the mixed number of good and bad texts, comments ,emotions, opinions, hatred etc., to collect the relevant tweet data there are many publicly used dataset providers such as one provided by Kaggle. The second step is to Preprocessing the extracted data by removing the extra words, comments, punctuations, special characters if needed stemming and lemmatization processes can be used. The feature extraction process can be used for the bagging of words with the combination of alphabet, numbers, special characters etc., Here comes the next step for the process of model selection, Choose the best deep learning or machine learning model for the job. For this task, recurrent neural networks (RNNs) that can manage sequential input and comprehend the context of the text, like Long Short-Term Memory (LSTM) or Gated Recurrent Units (GRUs), are a good option. Then after model selection , model training process is done here the process of training the data sets are done by feature extracted data, by using the techniques like cross validation and the hyperparameter tuning to optimize the model's performance. Then after model training the model evaluation process is done for the evaluating the accurate results of the data which come under the bullies criteria ,using that evaluated result the deployment process is done by calculating the total values of Training , Validation as accuracy and loss for measuring the models performance.

RNN's(Recurrent Neural Networks) algorithm and LSTM(Long Short-Term Memory) algorithms are the type of neural algorithms used for the process of visualizing data as the sequential data ,in this detection of bullies in tweets they are used for the detection of bullies in tweets by the text format and extractions and uses Bidirectional LSTM for the detection process before the algorithms used in the datasets they were text vectorized for the extraction and perfect detection evaluation process, where this process is done under the Natural Language Processing(NLP) , then RNN's were used for the comparing process of previously evaluated Training sets which acts as the self looping neural network algorithm and LSTM which is used for the vanishing of gradient problem with Recurrent Linear Activation Unit processing layer (relu) which takes place in the hidden layers in bidirectional LSTM Algorithm which regulates the flow of information in and out as in the form of gates, The transformation of independent activations into dependant ones is first carried out by the recurrent network. Moreover, it gives each layer the identical bias and weight, which lessens the complexity of the RNN's parameters. Additionally, it offers a consistent framework for remembering the previous outputs by using the previous output as an input to the following layer. (fig -1)[11]

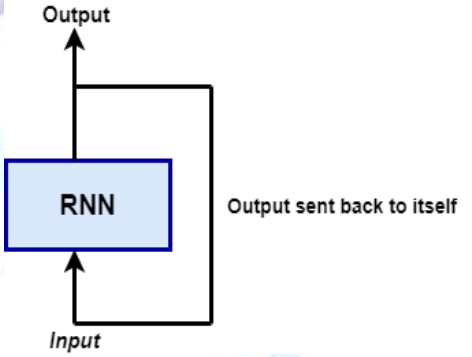


Fig-1 (RNN's Process)

Then the values were optimized by different epochs which means per episode training and the final outcomes where displayed as the Chart which is used for the visualization of the values as Training Accuracy, Validation Accuracy, Training Loss and Validation Loss Values as Numbers . If the bullies found where high in rate then the report can be sent to the consent who is responsible for the problem in the platform and To ensure that the model is effectively detecting cyberbullying, it is crucial to have a wide and diverse dataset for training it as well as to continuously check its performance.

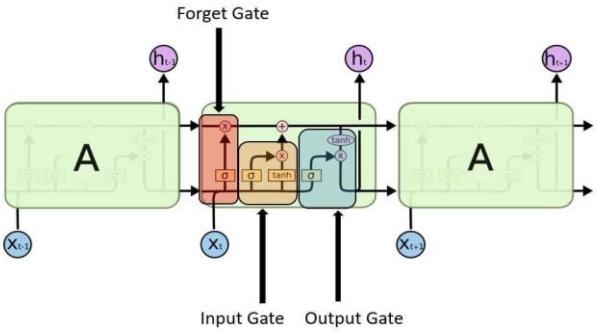
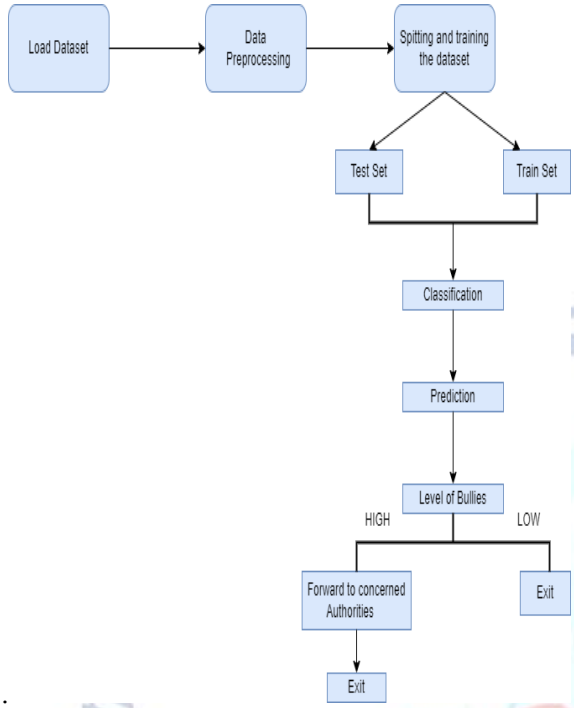


Fig-2 (LSTM)

The process of LSTM is given in the fig (2)[12] .Where there are there types of gates under LSTM process which where used for the Memory process with the sigmoid function determines whether to pass through values of 0 or 1. The tanh function also assigns weight to the values

that are passed, determining their importance on a scale from -1 to 1.The Fig(3) represents the flow of the process



.

Fig-3 (Data flow diagram)

4. RESULT AND ANALYSIS

The fig (4) represents the input of the data set which are extracted from the Twitter through the open platform for example Kaggle, the dataset is used for the detection of bullies and nonbullies with the label of 0,1 where zero represents the nonbullies ,1 represents the bullies. For this process at first the set of tweets are classified and vectorized using the process of sentimental analysis and the text vectorization process is used for the extraction of tweets similar to the feature of bullies criteria the fig(5,6) represents the above vectorization process which are converted into the form of array which are stored in the vectorization layer,

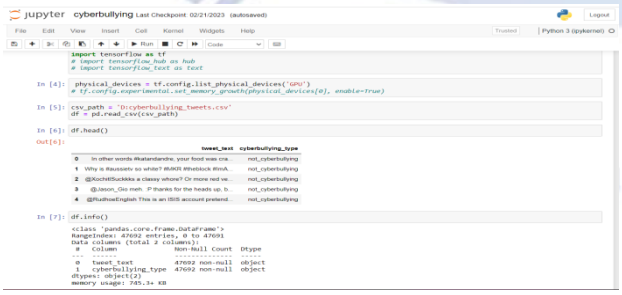


Fig - 4 (Data extracted for detection)

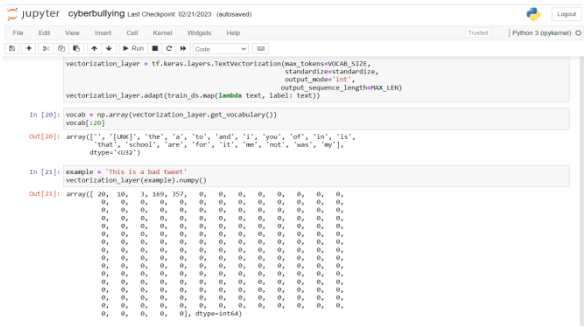


Fig – 5 (Text Vectorization process)

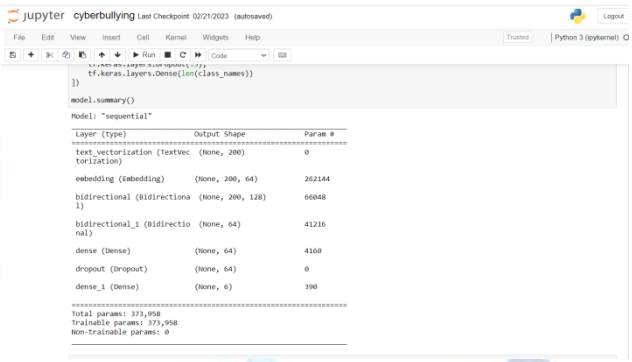
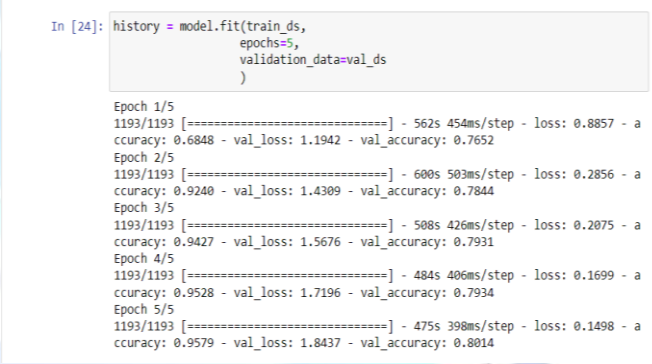


Fig - 6 (Text vectorization params)

The fig - 7 represents the running process of Epoch for the detection of bully tweets as the accuracy in training sets and validation sets, and the loss in training sets and validation sets .



```
In [24]: history = model.fit(train_ds,
                            epochs=5,
                            validation_data=val_ds
                            )

Epoch 1/5
1193/1193 [==============================] - 562s 454ms/step - loss: 0.8857 - a
ccuracy: 0.6848 - val_loss: 1.1942 - val_accuracy: 0.7652
Epoch 2/5
1193/1193 [==============================] - 600s 503ms/step - loss: 0.2856 - a
ccuracy: 0.9240 - val_loss: 1.4309 - val_accuracy: 0.7844
Epoch 3/5
1193/1193 [==============================] - 508s 426ms/step - loss: 0.2075 - a
ccuracy: 0.9427 - val_loss: 1.5676 - val_accuracy: 0.7931
Epoch 4/5
1193/1193 [==============================] - 484s 406ms/step - loss: 0.1699 - a
ccuracy: 0.9528 - val_loss: 1.7196 - val_accuracy: 0.7934
Epoch 5/5
1193/1193 [==============================] - 475s 398ms/step - loss: 0.1498 - a
ccuracy: 0.9579 - val_loss: 1.8437 - val_accuracy: 0.8014
```

Fig-7 (Running epoch process)

The Fig -8 represents the accuracy and loss of both training and validation set of bullies and nonbullies

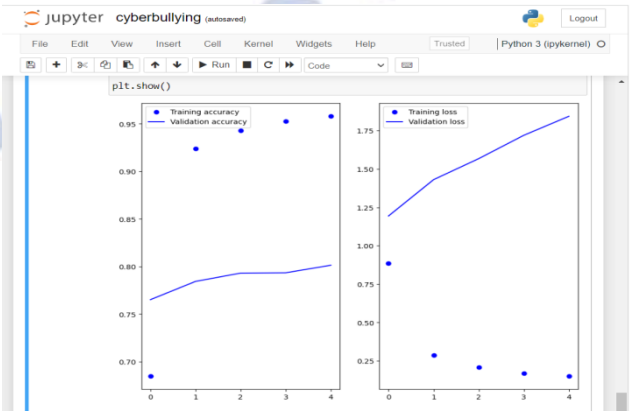


Fig -8 (Final Result of detection)

6. FUTURE SCOPE AND CONCLUSION

Cyberbullying on social media platforms can be identified and stopped using machine learning algorithms like RNN and LSTM. Large amounts of text data can be analysed by these algorithms to find patterns and words that are frequently used in cyberbullying activity. These algorithms can be taught to reliably recognise and indicate prospective cases of cyberbullying in real-time by being trained on a huge dataset of known cyberbullying incidents. This may enable social media platforms to avoid cyberbullying before it worsens by taking preventive actions. It is crucial to keep in mind, though, that machine learning algorithms are not perfect and occasionally fall short in their ability to spot cases of cyberbullying.Because of this, it's crucial to combine thesealgorithms with human moderators who can assess highlighted occurrences and reach decisions. In general, using machine learning techniques like RNN and LSTM can be a useful weapon in the struggle against cyberbullying on social media sites.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] V. Jain, V. Kumar, V. Pal and D. K. Vishwakarma, "Detection of Cyberbullying on Social Media Using Machine learning," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2021, pp. 1091-1096, doi: 10.1109/ICCMC51019.2021.9418254.

[2] M. Di Capua, E. Di Nardo and A. Petrosino, "Unsupervised cyber bullying detection in social networks," 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 2016, pp. 432-437, doi: 10.1109/ICPR.2016.7899672.

[3] J. Yadav, D. Kumar and D. Chauhan, "Cyberbullying Detection using Pre-Trained BERT Model," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2020, pp. 1096-1100, doi: 10.1109/ICESC48915.2020.9155700.

[4] R. R. Dalvi, S. Baliram Chavan and A. Halbe, "Detecting A Twitter Cyberbullying Using Machine Learning," 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2020, pp. 297-301, doi: 10.1109/ICICCS48265.2020.9120893.

[5] Trana, R.E., Gomez, C.E., Adler, R.F. (2021). Fighting Cyberbullying: An Analysis of Algorithms Used to Detect Harassing Text Found on YouTube. In: Ahram, T. (eds) Advances in Artificial Intelligence, Software and Systems Engineering. AHFE 2020. Advances in Intelligent Systems and Computing, vol 1213. Springer, Cham. https://doi.org/10.1007/978-3-030-51328-3_2

[6] N. Tsapatsoulis and V. Anastasopoulou, "Cyberbullies in Twitter: A focused review," 2019 14th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), Larnaca, Cyprus, 2019, pp. 1-6, doi: 10.1109/SMAP.2019.8864918.

[7] G. A. León-Paredes et al., "Presumptive Detection of Cyberbullying on Twitter through Natural Language Processing and Machine Learning in the Spanish Language," 2019 IEEE CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON), Valparaiso, Chile, 2019, pp. 1-7, doi: 10.1109/CHILECON47746.2019.8987684.

[8] P. K. Roy, A. K. Tripathy, T. K. Das and X. -Z. Gao, "A Framework for Hate Speech Detection Using Deep Convolutional Neural Network," in IEEE Access, vol. 8, pp. 204951-204962, 2020, doi: 10.1109/ACCESS.2020.3037073.

[9] S. M. Kargutkar and V. Chitre, "A Study of Cyberbullying Detection Using Machine Learning Techniques," 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2020, pp. 734-739, doi: 10.1109/ICCMC48092.2020.ICCMC-000137.

[10] https://www.javatpoint.com/recurrent-neural-network-in-tensorflow#:~:text=A%20recurrent%20neural%20network%20(RNN,neurons%20in%20the%20human%20brain

[11] https://www.javatpoint.com/long-short-term-memory-rnn-in-tensorflow

[12] https://www.itm-conferences.org/articles/itmconf/pdf/2021/05/itmconf_icacc2021_03038.pdf