iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VII **Month of publication:** July 2022

DOI: https://doi.org/10.22214/ijraset.2022.45872

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

Sentiment Analysis on Covid-19 Using Deep Learning

Soni Mehta[1], Shruti Pednekar[2]
[1]*Student,* [2]*Project Guide, Assistant Professor, Department of Computer Science and Engineering, GEC, Farmagudi - GOA*

Abstract: ***The social media has immense popularity among all the services today. Twitter is one of the widely used platform by people to express their opinions and display sentiments on different occasions. Sentiment Analysis is a classification task in order to identify public reviews about different issues like reviews about movies, restaurants and other current issues by extracting the public reviews from social media. As we all know the world was hit by a global pandemic of COVID-19 and it is still a global issue. Due to rapid increase in the infection people were expressing their emotions, thoughts, and were having mixed feelings regarding the situation. The main objective of this research paper is to analyze the emotions expressed by people using twitter data. The corona specific tweets are collected from twitter and pre-processing is done to clean the data and later word embedding pre-trained model is used and finally CNN, LSTM and CNN-BiLSTM hybrid deep learning approaches are applied. The mode l is evaluated using accuracy, precision and recall techniques.***
Keywords: sentiment analysis, twitter data, covid-19, word embedding, deep learning.

I. INTRODUCTION

The rise of Internet technology has played an unprecedented role in increasing the number of social media and e-commerce platforms. In addition, users are now accustomed to the idea of expressing their feelings and emotions with others by using these platforms either by text or multimedia data. This phenomenon has resulted in the production and generation of a large variety of data, which can be analysed for assessing sentiment. It is beneficial for individuals and organizations to analyze sentiment, especially given this immense production of data.

Sentiment Analysis has attracted a lot of attention from researches from recent times. Nowadays, a massive amount of information exchange through web-based technologies and Internet-based users are expressing their emotions through social media, microblogging sites, and other media with the help of the Internet. The Sentiment Analysis is the most important methodology to classify the user's opinions, emotions to determine whether the particular outlook is positive, negative, or neutral feedback on certain issues like movie reviews, political opinions, global pandemics, and other economic crises.

Deep learning involves applying artificial neural networks to learn different tasks using networks that are attributed to different layers. The search primarily takes inspiration from the way that the human brain is structured, as it contains a large number of entities (neurons) that are used for processing the information. The use of neural networks plays an important role at different levels for analyzing sentiment, including the document level, aspect level, and sentence level.

In this paper, we have proposed CNN, LSTM and CNN-BiLSTM method for the sentiment analysis of twitter data having 1.7 lakh tweets. The approach mainly consists of extracting the tweets, pre-processing of the tweets, applying word embedding model to extract word embedding for words using Glove and FastText pre-trained models and finally comparing the deep learning approaches.

II. RELATED WORK

Sentiment analysis involves investigating the approach of a writer toward a particular subject or the overall contextual polarity of an entire document. With the major increment in the amount of online data generated, Numerous studies have used sentiment analysis to classify texts based on sentiment or opinion using various machine learning and deep learning approaches. In recent times, sentiment analysis has involved a substantial amount of work and is growing rapidly.

Akshat Shrivastava [2] this paper utilized the live twitter dataset where the pre-processor is applied to the crude sentences. Further, the diverse ML strategies prepare the dataset with highlights and afterward the semantic investigation offers an enormous arrangement of equivalent words and comparability which gives the extremity of the substance. Naïve bayes, maximum entropy and support vector machine algorithm is used to classify data.

Sani Kamis [3], This study presents a comparison of different deep learning methods used for sentiment analysis in Twitter data.

Particularly, two categories of neural networks are utilized, convolutional neural networks (CNN) and recurrent neural networks (RNN). Additionally, different word embedding systems such as the Word2Vec and the global vectors for word representation (GloVe) models are compared. Various tests and combinations are applied and best scoring values for each model are compared in terms of their performance.

Manoj Sethi [4] in this study, the sole focus is to analyze the emotions expressed by people using social media such as Twitter etc. Corona specific tweets are acquired from twitter platform. After gathering the tweets, they are labelled and a model is developed which is effective for detecting the actual sentiment behind a tweet related to COVID-19. The substantial assessments are performed in bi-class and multi-class setting over n-gram feature set along with cross-dataset evaluation of different machine learning techniques like logistic regression, XG boost, SVM in order to develop the model.

Nikhil Yadav[5], This paper emphasizes the different techniques utilized for classifying the product critiques (which can be within the form of tweets) according to critiques expressed in tweets to analyze whether or not the massive behaviour is positive, negative or neutral and use of that analysis for the evaluation of product market. Data used in this look at our online product critiques gathered from twitter and used to rank the satisfactory classifier for sentiments.

III. METHODOLOGY

A. Dataset Description

Twitter is used as the primary data source in order to gather tweets specific to corona virus. The data used in this work is comprised of a large number of tweets collected from GitHub repository [1]. The collected data is in form of positive, negative and neutral tweets. We have total of 1.79 lakh tweets. The dataset has 7 attributes. The attribute details are shown in Table I and Table II.

```
data.shape
```

```
(179108, 7)
```

Table I

	user_name	user_location	user_description	user_followers	text	hashtags	source
0	á□□á□¥â□»Õ¬é□□ Ì®	astroworld	wednesday addams as a disney princess keepin i...	624	If I smelled the scent of hand sanitizers toda...	NaN	Twitter for iPhone
1	Tom Basile ð□□ºð□□¸	New York, NY	Husband, Father, Columnist & Commentator. Auth...	2253	Hey @Yankees @YankeesPR and @MLB - wouldn't it...	NaN	Twitter for Android
2	Time4fisticuffs	Pewee Valley, KY	#Christian #Catholic #Conservative #Reagan #Re...	9275	@diane3443 @wdunlap @realDonaldTrump Trump nev...	['COVID19']	Twitter for Android
3	ethel mertz	Stuck in the Middle	#Browns #Indians #ClevelandProud #[]_[] #Cavs ...	197	@brookbanktv The one gift #COVID19 has give me...	['COVID19']	Twitter for iPhone
4	DIPR-J&K	Jammu and Kashmir	ð□□□ï¸□Official Twitter handle of Department o...	101009	25 July : Media Bulletin on Novel #CoronaVirus...	['CoronaVirusUpdates', 'COVID19']	Twitter for Android

Table II

	user_name	user_location	user_description	user_followers	text	hashtags	source
179103	AJIMATI AbdulRahman O.	Ilorin, Nigeria	Animal Scientist|| Muslim|| Real Madrid/Chelsea	412	Thanks @IamOhmai for nominating me for the @WH...	['WearAMask']	Twitter for Android
179104	Jason	Ontario	When your cat has more baking soda than Ninja ...	150	2020! The year of insanity! Lol! #COVID19 http...	['COVID19']	Twitter for Android
179105	BEEHEMOTH â□³	ð□□¨ð□□¦ Canada	â□□ï¸□ The Architects of Free Trade â□□ï¸□ Rea...	1623	@CTVNews A powerful painting by Juan Lucena. I...	NaN	Twitter Web App
179106	Gary DelPonte	New York City	Global UX UI Visual Designer. StoryTeller, Mus...	1338	More than 1,200 students test positive for #CO...	['COVID19']	Twitter for iPhone
179107	TUKY II	Aliwal North, South Africa	TOKELO SEKHOPA | TUKY II | LAST BORN | EISH TU...	97	I stop when I see a Stop\n\n@SABCNews\n@Izinda...	NaN	Twitter for Android

B. Pre-Processing Dataset

Raw tweets scratched from the twitter contain lots of noise data, misspelt words, include various abbreviations, emojis and so on this is because of the easy-going ideas of individuals utilization of social media. These words often disturb the sentiments of the tweets and might not give perfect accuracy. Therefore, tweets are pre-processed before performing word embedding techniques. Pre-processing basically means cleaning and removing of non-textual substances from the dataset to improve the performance of the proposed model. Below are the steps used for pre-processing of tweets:

1) Converting upper case to lower case.
2) Strip spaces and quotes ("and') from the ends of tweet.
3) Removal of re-tweets, URL, Hashtags, Mentions.
4) Stopword Removal: Stop words that don't affect the meaning of the text or the sentences are removed for example and, or, still, the, which etc…
5) Tokenization and lemmatization: Tokenization is method of separating the text into the words which are called tokens and lemmatization are used to break these words into typical root word for instances, the word "troubling" "troubled" "troubles" will be diminished into a root word "trouble".

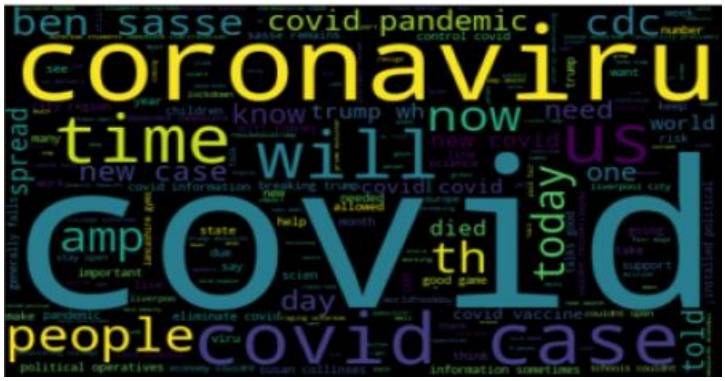


Fig. 1 Word Cloud

C. Word Embedding

Word embedding is a method of a feature extraction which will help in boosting the accuracy of deep learning, machine learning and natural language processing task by converting each word into its corresponding vector represented using pre-trained model and our dataset.

Feature extraction in the case of sentiment analysis is used to analyze sentiments from subjective texts and to find the polarity by classifying the data into positive, negative, and neutral. Table III shows us the cleaned data after the pre-processing is done and feature extraction is applied.

Table III

	text	Subjectivity	Polarity	Analysis	label	words	sentences	lemmatizer_tweets	clean_tweets
0	smelled scent hand sanitizers today someone pa...	0.250000	-0.25	Negative	0	[smelled, scent, hand, sanitizers, today, some...	smelled scent hand sanitizers today someone pa...	[smelled, scent, hand, sanitizers, today, some...	smelled scent hand sanitizers today someone pa...
1	hey wouldnt made sense players pay respects	0.000000	0.00	Neutral	1	[hey, wouldnt, made, sense, players, pay, resp...	hey wouldnt made sense players pay respects	[hey, wouldnt, made, sense, player, pay, respect]	hey wouldnt made sense players pay respects
2	trump never claimed covid19 hoax claim effort	0.000000	0.00	Neutral	1	[trump, never, claimed, hoax, claim, effort]	trump never claimed covid19 hoax claim effort	[trump, never, claimed, covid19, hoax, claim, ...	trump never claimed covid19 hoax claim effort
3	one gift covid19 give appreciation simple thin...	0.357143	0.00	Neutral	1	[one, gift, give, appreciation, simple, things...	one gift covid19 give appreciation simple thin...	[one, gift, covid19, give, appreciation, simpl...	one gift covid19 give appreciation simple thin...
4	25 july media bulletin novel coronavirusupdate...	0.000000	0.00	Neutral	1	[july, media, bulletin, novel, coronavirusupda...	25 july media bulletin novel coronavirusupdate...	[25, july, medium, bulletin, novel, coronaviru...	25 july media bulletin novel coronavirusupdate...

The two basic pre-trained models we will be using are FastText and Glove(Global Vectors for word representation).

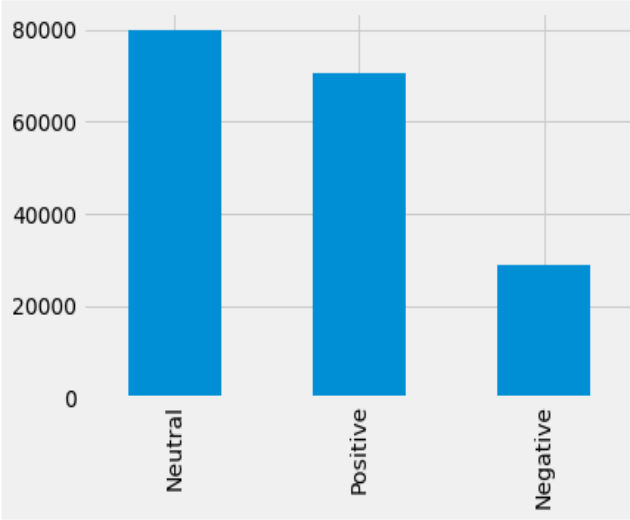
1) Glove: GloVe is an unsupervised learning algorithm to learn vector representation i.e word embedding for various words. The model was proposed by Pennington et al. (2014). The GloVe is a pre-trained word embedding method used for solving deep learning. It is a modified version of word2Vec embedding techniques in order to solve the problem of word2Vec weakness. We have used the 6B version of the GloVe vector with 200 billion tokenized web data with 200-dimensional wordvectors in this project.

2) FastText: FastText is an NLP library developed by the Facebook research team for text classification and word embeddings. FastText is popular due to its training speed and accuracy. There are two frameworks of FastText Text Representation (fastText word embeddings)and Text Classification. In FastText each word is represented as a bag of character n-gram. FastText word embeddings supports both Continuous Bag of Words (CBOW) and Skip-Gram models For this project, we will use Gensim fastText library to train fastText word embeddings in Python.

```
loading word embeddings...
999995it [09:52, 1688.81it/s]
found 999995 word vectors
```

Fig. 2 Loading FastText

Table IV



D. Deep Learning Approaches

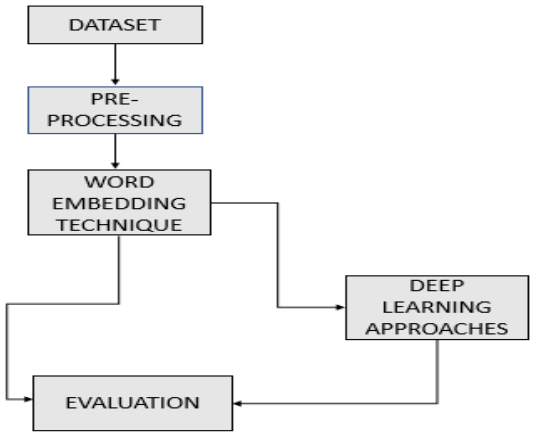


Fig.3 Proposed Architecture

E. Convolution Neural Network (CNN)

We used keras with TensorFlow backend to implement the Convolutional Neural Network model. We used the dense vector representation of the tweets to train our CNN models. The function of CNN is to extract effective features from sequences. The first step is using the word vector matrix as the input to CNN model; the second step is to use convolution kernel to construct local n-gram features from word vector matrix; the third step is the maximum pool of the result of each convolution kernel; We perform temporal convolution with a kernel size of 64 and zero padding. After the convolution layer, we apply relu activation function and then perform Max Pooling over time to reduce the dimensionality of the data. We also added dense and dropout layers after the embedding layer and the fully connected layer to regularize our network and prevent it from overfitting-Pooling is a way of feature processing in convolution neural network, usually after convolution operation. The purpose of pooling is to calculate the local sufficient statistics. Max-pooling can extract the most important features from the convolution layer.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 400, 200)          4000000

 conv1d (Conv1D)             (None, 398, 250)          150250

 max_pooling1d (MaxPooling1D  (None, 199, 250)         0
 )

 conv1d_1 (Conv1D)           (None, 195, 250)          312750

 global_max_pooling1d (Globa  (None, 250)              0
 lMaxPooling1D)

 dense (Dense)               (None, 250)               62750

 dropout (Dropout)           (None, 250)               0

 dense_1 (Dense)             (None, 1)                 251

=================================================================
Total params: 4,526,001
Trainable params: 526,001
Non-trainable params: 4,000,000
_________________________________________________________________
```

Fig.4 Details of CNN Model

F. Long-Short Term Memory (LSTM)

Next, we use LSTM model for training our dataset. Here, the embedding layer is followed by an LSTM layer where we experimented with different number of LSTM units. The LSTM layer is followed by a fully-connected layer with 128 units and relu activation. In this configuration a single LSTM layer is used with a dropout of 60%. LSTM consists of two states: hidden state and cell state. At a particular time, step t, LSTM decides which information must be taken from the state of the cell. The decision is made by a sigmoid function layer called the forget gate.

```
Model: "model"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_1 (InputLayer)        [(None, 400)]             0

 embedding_1 (Embedding)     (None, 400, 200)          4000000

 lstm_layer (LSTM)           (None, 400, 128)          168448

 global_max_pooling1d_1 (Glo  (None, 128)              0
 balMaxPooling1D)

 dropout_1 (Dropout)         (None, 128)               0

 dense_2 (Dense)             (None, 90)                11610

 dropout_2 (Dropout)         (None, 90)                0

 dense_3 (Dense)             (None, 60)                5460

 dropout_3 (Dropout)         (None, 60)                0

 dense_4 (Dense)             (None, 1)                 61

=================================================================
Total params: 4,185,579
Trainable params: 4,185,579
Non-trainable params: 0
_________________________________________________________________
```

Fig.5 Details of LSTM Model

G. Convolution Neural Network with Bidirectional Long-Short Term Memory (CNN-BiLSTM)

In CNN-BiLSTM approach, First embedding layer passes features into drop out layer with the rate to 0.2 to avoid over fitting. The output feds into first 1-Dimensional CNN layer. Further, we will define 64 filters. This allows us to train 64 different features on the second layer of the network. The result will be fed into the Bi-LSTM layer of size 64 to capture long range dependencies to extract feature and then feed into the Fully Connected Dense layer. Next, the output of Dense Layer passes to drop out with a rate of 256, the purpose of which to drop some randomly weights of the matrix. Finally, we get the fully connected dense layer with sigmoid function.

```
Model: "sequential_11"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_11 (Embedding)    (None, 400, 200)          4000000

 dropout_24 (Dropout)        (None, 400, 200)          0

 conv1d_8 (Conv1D)           (None, 398, 64)           38464

 bidirectional_5 (Bidirectio  (None, 512)              657408
 nal)

 dense_11 (Dense)            (None, 256)               131328

 dropout_25 (Dropout)        (None, 256)               0

 dense_12 (Dense)            (None, 256)               65792

 dense_13 (Dense)            (None, 1)                 257

=================================================================
Total params: 4,893,249
Trainable params: 893,249
Non-trainable params: 4,000,000
_________________________________________________________________
```

Fig.6 Details of CNN-BiLSTM model

Embedding Word System: FastText	
MODELS	ACCURACY (%)
CNN	52.01%
LSTM	51.23%
CNN-BiLSTM	60.55%

Embedding Word System: Glove	
MODELS	ACCURACY (%)
CNN	46.86%
LSTM	44.54%
CNN-BiLSTM	58.11%

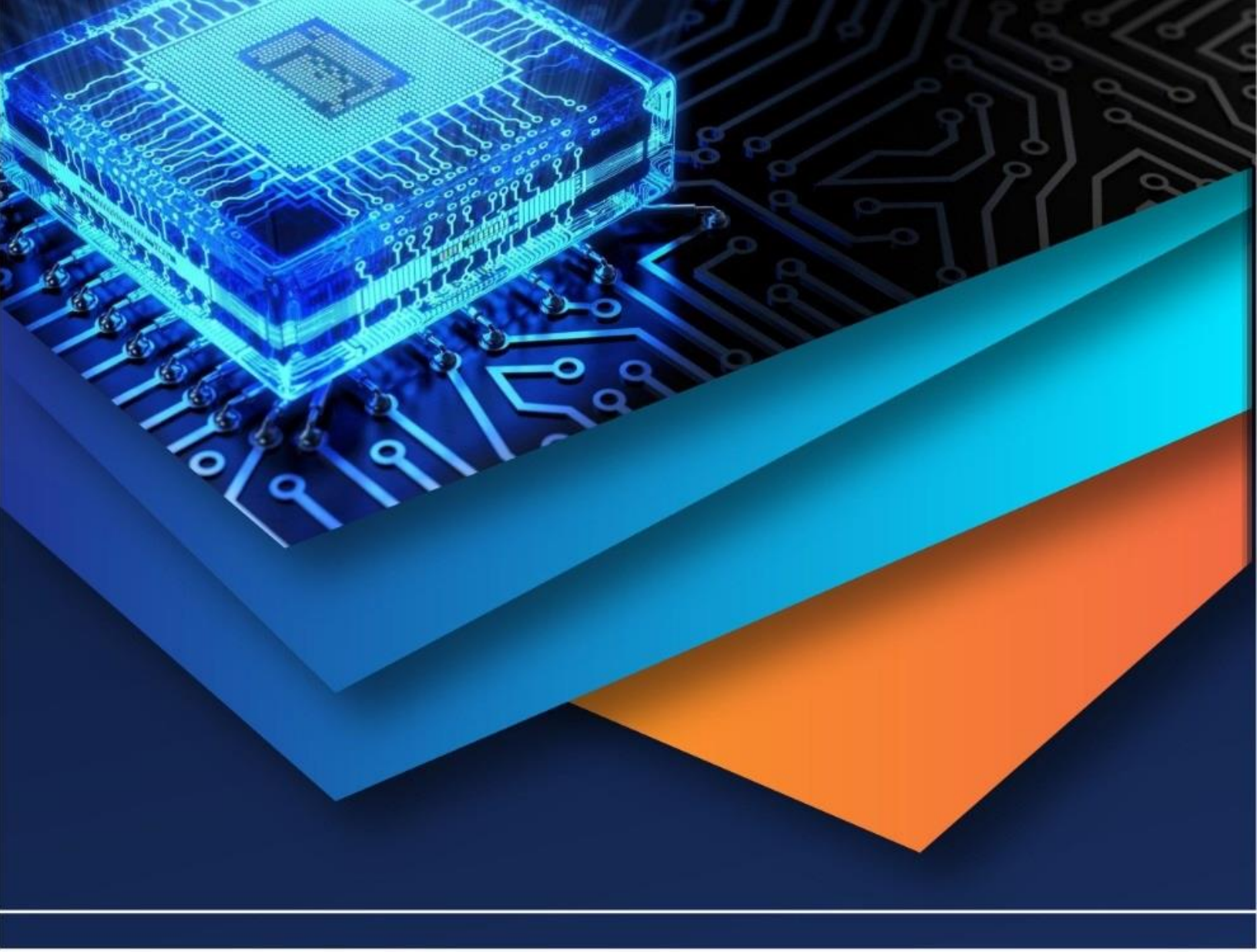
IV. CONCLUSION

Sentiment Analysis is a task to realize the expression of public feelings expressed on social media about different issues to detect whether the people's outlook is positive, negative, and neutral. The main objective of this project was to detect and classify the public emotion, which is talked about the COVID-19 global pandemic using a deep learning (CNN, LSTM, CNN-BiLSTM) model with the help of open sources pre-trained (FastText and GloVe) models in order to get good accuracy. The Overall experiments shows 58% accuracy of CNN-BiLSTM model obtained using Glove pre-trained model and 60% accuracy obtained using FastText pre-trained model of CNN-BiLSTM model. It is seen that the hybrid model provides better result and accuracy than the single CNN and LSTM approach.

In future work, we can try to extract more dataset and apply more hybrid deep learning approaches to get better accuracy and better classification results.

REFERENCES

[1] https://www.kaggle.com/gpreda/covid19-tweets.
[2] Akshat Shrivastava, Anurag Sen, Amritansh Shrivastava, Sachin Singh, Nagesh Jadhav, “Collective Intelligence Sentiment Analysis of Tweets using Machine Learning”, International Journal of Scientific & Engineering Research Volume 11, Issue 6, June-2020.
[3] Sani Kamis , Dionysis Goularas, “Evaluation of Deep Learning Techniques in Sentiment Analysis from Twitter Data”, 2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML).
[4] Manoj Sethi , Sarthak Pandey , Prashant Trar, Prateek Soni, “Sentiment Identification in COVID-19 Specific Tweets”, Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC 2020).
[5] Nikhil Yadav, Omkar Kudale, Srishti Gupta, Aditi Rao, "Twitter Sentiment Analysis Using Machine Learning For Product Evaluation",Proceedings of the Fifth International Conference on Inventive Computation Technologies (ICICT-2020).
[6] Vishu Tyagi, Ashwini Kumar, Sanjoy Das, “Sentiment Analysis on Twitter Data Using Deep Learning approach”, 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN).
[7] Anu J Nair, Veena G, Aadithya Vinayak,"Comparative study of Twitter Sentiment On COVID - 19 Tweets",Proceedings of the Fifth International Conference on Computing Methodologies and Communication (ICCMC 2021).
[8] Bhumika Gupta, Monika Negi, Kanika Vishwakarma, Goldi Rawat, Priyanka Badhani, "Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python",,International Journal of Computer Applications (0975 – 8887)Volume 165 – No.9, May 2017.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY