iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** X **Month of publication:** October 2024

DOI: https://doi.org/10.22214/ijraset.2024.64700

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 12 Issue X Oct 2024- Available at www.ijraset.com

Supervised Sentiment Analysis of COVID-19 Tweets: Insights into Public Reactions and Trends

Umesha D K[1], Priya M K[2], Nandan N[3]
[1]*Lecturer, Department of Computer Science & Engineering, Government Polytechnic, Sorab, Shimoga-577429, Karnataka, India*
[2]*Lecturer, Department of Computer Science & Engineering, Government Polytechnic, Ramanagara-562159, Karnataka, India*
[3]*Lecturer, Department of Computer Science & Engineering, Government Polytechnic, Arakere, Srirangapattana-571434, Karnataka, India*

Abstract: ***In the modern digital era, social media platforms have experienced exponential growth, serving as a medium for individuals to share their emotions and opinions. Twitter, a prominent social media platform, hosts a substantial number of users who regularly express their thoughts through tweets. This research paper focuses on conducting sentiment analysis on tweets pertaining to the COVID-19 situation in India. By utilizing a lexicon-based approach using TextBlob, we determine sentiment scores for each tweet. These scores are then used to categorize the tweets into Positive, Negative, or Neutral sentiment categories. To further enhance the accuracy of sentiment classification, we employ the Naïve Bayes machine learning model, achieving an impressive accuracy rate of 87.65%. The sentiment analysis results obtained through this study provide valuable insights for governmental bodies and other organizations, enabling them to formulate effective policies in order to address similar issues in the future.***

Keywords: Sentimental Analysis, Machine Learning, Naive Bayes, Tweets, TextBlob, Lexicon-based approach.

I. INTRODUCTION

Corona virus disease 2019 (COVID-19) is an infectious disease caused by Severe Acute Respiratory Syndrome Corona Virus 2 (SARS-CoV-2). It was first identified in December 2019 in Wuhan, Hubei, China, and has resulted in an ongoing pandemic [1]. People are posting so many tweets related to COVID-19 in the Twitter to express their emotions and feelings. To control the current outbreak there is a need for statistical and quantitative analysis of people emotions which are expressed in the tweets [2].

Sentiment analysis is one such technique which is used to analyse the opinion, feelings and emotions in the social media posts. Sentiment analysis tools can be effectively used to gauge public sentiment trends from social media data, especially given the high levels of trust in social media posts among many networks **[3].** The most common type of sentiment analysis is called 'polarity detection' and involves classifying a statement as 'positive', 'negative', or 'neutral'.

This research has been done to study the impact of COVID-19 pandemic on the Indian people's mind by analysing the sentiments of their comments expressed on social media platforms like Twitter. *TextBlob* [4] was used for calculating the sentiment scores of tweets and then tweets were classified as 'Positive', 'Negative' or 'Neutral' category. Finally Naive Bayes machine learning [9] model was trained using these classified tweets.

II. LITERATURE SURVEY

Many researchers are working on Sentiment Analysis on Twitter data, with the major increment in the amount of data generated online; the field has attracted a significant number of researchers to invest in the study of social media data and to make the most out of the information available.

A. P. Jain et al [5] provided a detailed approach for sentiment analysis using Machine Learning approach. Sentiment analysis is done by applying multinomial naive bayes and decision tree models in the proposed method.

Manoj Sethi et al[6], done COVID-19 Twitter data analysis using bi-class and multi-class classification. They have compared Decision Tree, SVM and Random Forest classification models.

Nithyashree T and Nirmala M.B [7] worked on Sentiment analysis of Hotels review tweets. The SVM machine learning technique is used for classification which labelled the tweets as positive with accuracy 61.12%.

M. Rathi, et al [8], done research on the classification of emotions of different kind of tweets' data using SVM, ADABOOSTED DECISION TREE and DECISION TREE based hybrid sentiment classification model. Hybrid model using ensemble technic gave more accuracy than individual technique

P. P. Surya and B. Subbulakshmi[9], have proposed a work in which twitter data related to COVID -19 is analysed. COVID-19 tweets were classified using lexicon-based method into basic emotions like anger, anticipation, disgust, fear, joy, sadness, surprise and trust.

III. PROPOSED METHOD

In this research we have used sentimental analysis of tweets related to COVID-19 in India. Twitter data is collected and pre-processed to clean the data. After cleaning, data is analysed and classified using TextBlob library. Steps in the proposed method are shown in Figure-1.

Figure 1: Steps in Proposed Method

A. Data Collection

The data set was collected from the website "https://ieee-dataport.org/open-access/coronavirus-covid-19-tweets-dataset" [10]. Portion of the data set is shown in Table-1. The dataset contains 648958 tweets related to COVID-19 from 20/03/2020 to 31/05/2020 and each tweet has four columns which are described as follows.

1) Text_Id: It contains unique ID for each tweet.
2) Tweet: It is the tweet text of that particular tweet ID.
3) Date: The date on which the tweet was tweeted.
4) Location: The place from where the tweet was tweeted.

Table 1: Portion of the Data set

	Text_Id	tweet	Date	Location
648953	1.266960e+18	For the first time perhaps, someone to took ti...	Sun May 31 04:52:36 +0000 2020	India
648954	1.266960e+18	RT @hvgoenka: 70 year Kamalamma offered an NGO...	Sun May 31 04:52:38 +0000 2020	Nagpur, India
648955	1.266960e+18	Odisha reports 129 new #COVID19 positive cases...	Sun May 31 04:52:40 +0000 2020	New Delhi, India
648956	1.266960e+18	@nidhiindiatv Happy rainy day ..\n\nHope rain ...	Sun May 31 04:52:40 +0000 2020	Muzaffarnagar, India
648957	1.266960e+18	RT @nsui: Kerala Student Union activists did S...	Sun May 31 04:52:40 +0000 2020	New Delhi, India

B. Data Pre-processing

Raw tweets scraped from twitter generally result in a noisy dataset. This is due to the casual nature of people's usage of social media. Tweets have certain special characteristics such as retweets, emoticons, user mentions, etc. which have to be suitably extracted. Therefore, raw twitter data has to be normalized to create a dataset which can be easily learned by various classifiers. We have applied an extensive number of pre-processing steps to standardize the dataset and reduce its size. We first do some general pre-processing on tweets which is as follows.

1) Converting all uppercase letters to lowercase.
2) Tokenization: It generally done by installing the NLP package. In this step we remove hash tags, numbers, URL's and user mentions (@).
3) Emoticon replacements: Emoticons are very important in determining the sentiment. So, the emoticons are replaced by their polarity by seeing the emoticon dictionary.
4) Removal of stop words: Stop words play a negative role in sentimental analysis, so it is important to be removed. They occur both in negative and positive tweets. A list of stop words like he, she, at, on, a, the, etc. are created and ignored.
5) Removal of Punctuations: In this step we remove any unwanted characters and punctuations which are not necessary for sentiment analysis.

C. Analysis of Tweets

We performed an analytical study on the constructed dataset in order to mine knowledge allowing an understanding and a mastery of the tweeters insights before performing data mining tasks, which help to provide a more comprehensive perception. The word cloud for 50 top words is shown in Figure-2. This figure shows that most of words which are used in the tweets. Figure-3 shows the frequency of top 25 words. We can remark that these words describe all the situation people are overcoming in real life during this hard period.

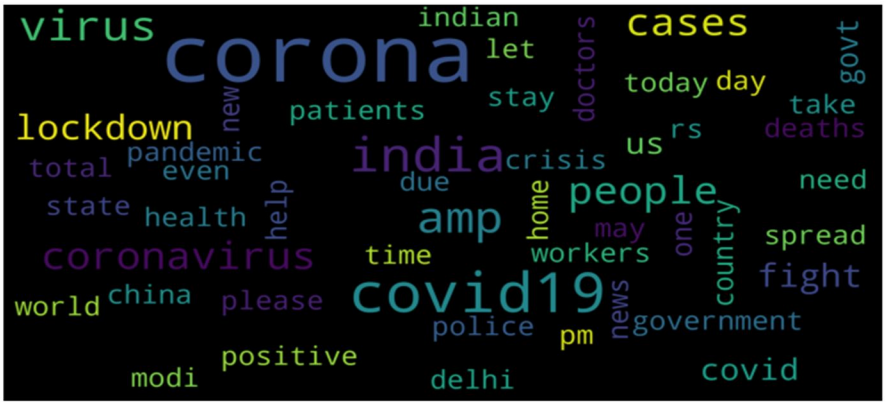


Figure 2. Word Cloud shows top 50 words.

D. Sentiments Analysis

The proposed method used machine learning based TextBlob library present in python for sentimental analysis [11]. After cleaning the dataset, to perform Sentiment analysis TextBlob library has been called. TextBlob is a Python library for processing textual data. It provides a simple API for diving into common Natural Language Processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation and more. The *sentiment* function of TextBlob returns two properties, polarity, and subjectivity [4].

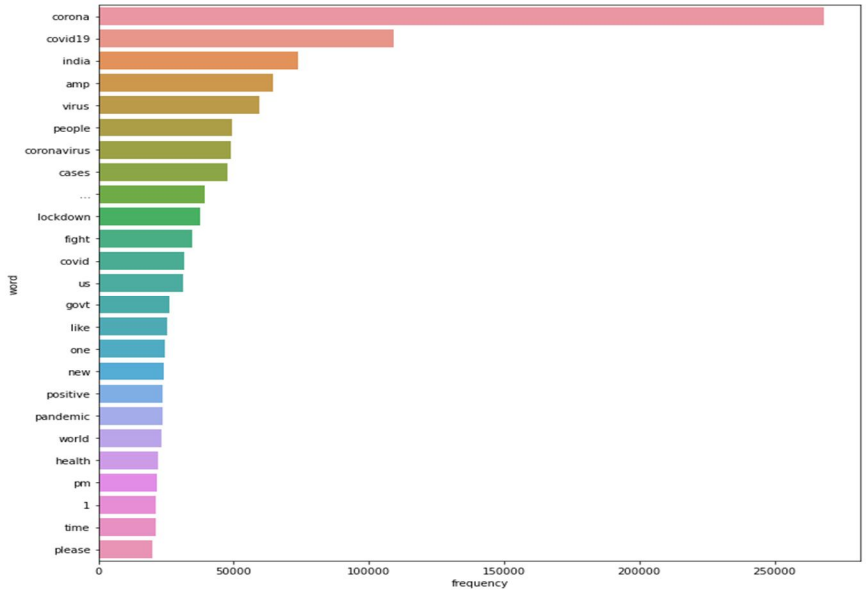


Figure 3. Frequency of top 25 words in the dataset

Polarity: It means emotions expressed in a sentence. Emotions are closely related to sentiments. It is a float which lies in the range of [-1,1] where

- Polarity Score < 0, Negative sentiment
- Polarity Score = 0, Neutral sentiment
- Polarity Score >0, Positive sentiment [6]

Subjectivity: Subjective sentences generally refer to personal opinion, emotion or judgment whereas objective refers to factual information. Subjectivity is also a float which lies in the range of [0,1].

```
from textblob import TextBlob

TextBlob("not a very great calculation").sentiment
## Sentiment(polarity=-0.3076923076923077, subjectivity=0.5769230769230769)
```

From the above example, it can be observed that words like "not a very great calculation" has a polarity =-0.3076 (<0) so it is a negative sentiment

```
TextBlob("very great").sentiment
## Sentiment(polarity=1.0, subjectivity=0.9750000000000001)
```

From the above example, it can be observed that the sentence "very great" has a polarity =1.0 (>0) so it is a positive sentiment.
Using the TextBlob package we have classified the tweets in the data set as "Negative", "Positive"And "Neutral" using polarity score. The Table 2 shown below describe the portion of classified data set.

Table-2: Portion of Dataset Classified using Polarity Score

	Text_Id	tweet	Date	Location	subjectivity	polarity	sentiment
648953	1.266960e+18	first time perhaps someone took time acknowled...	Sun May 31 04:52:36 +0000 2020	India	0.477778	0.388095	positive
648954	1.266960e+18	70 year kamalamma offered ngo bangalore rs 500...	Sun May 31 04:52:38 +0000 2020	Nagpur, India	0.400000	-0.250000	negative
648955	1.266960e+18	odisha reports 129 new covid19 positive cases ...	Sun May 31 04:52:40 +0000 2020	New Delhi, India	0.587500	0.057576	positive
648956	1.266960e+18	happy rainy day hope rain wash negative vibes ...	Sun May 31 04:52:40 +0000 2020	Muzaffarnagar, India	0.533333	0.200000	positive
648957	1.266960e+18	kerala student union activists sanitization sa...	Sun May 31 04:52:40 +0000 2020	New Delhi, India	0.000000	0.000000	neutral

E. Building Classifier Model.

Once all the tweets in the data set are labelled as "Positive"," Negative" and "Neutral" category using TextBlob library, next step is to train Naive bayes model using these tweets. Proposed method used Multinomial Naive Bayes algorithm to build the classifier model. Model is trained with 80% training data and validated using 20% test data.

IV. RESULTS AND DISCUSSION

Data set contains total 6,48,958 tweets related to COVID-19 and these tweets were classified into basic emotion categories. As illustrated in Table-3, it is observed that 2,82,514 (44%) of the tweets were classified as Neutral, 2,51,840 (39%) tweets were classified as Positive and 1,14,604(18%) tweets were classified as Negative tweets. This analysis shows that people were neutral about the COVID-19 situation. Figure-4 shows the sentimental analysis bar chart in which Neutral tweets are more in the data set. Figure-5 shows the distribution of tweets according polarity scores. It is observed from the figure that the tweets are more distributed towards the centre (Neutral)

Table -3

Categories	Total Number of Tweets	% of Tweets
Neutral	282514	44%
Positive	251840	39%
Negative	114604	18%

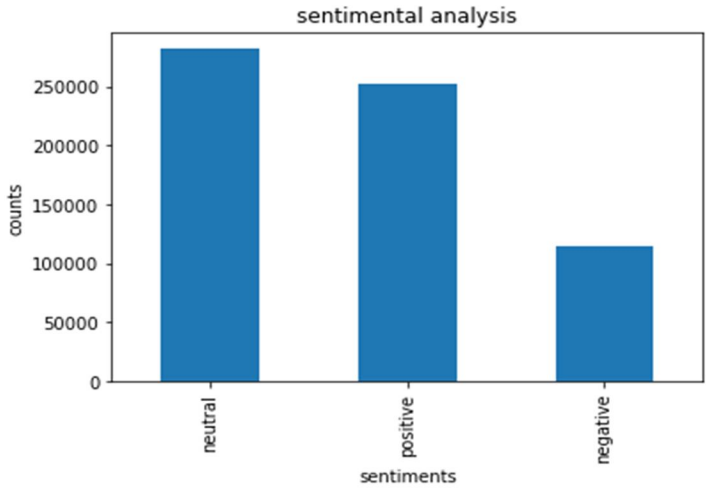


Figure 4: Classification of tweets into Sentiments

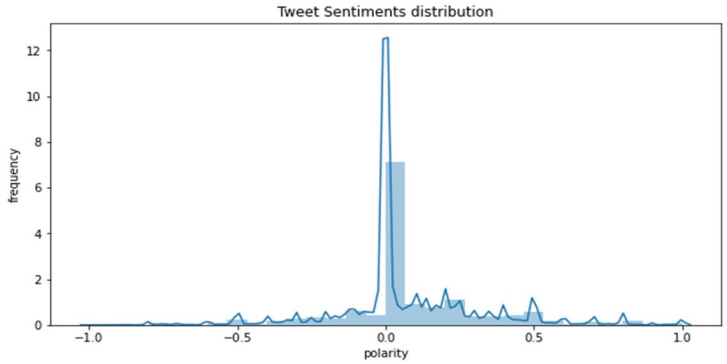


Figure 5: Tweets Distribution

Naïve bayes classifier was evaluated using accuracy as the metrics. For the proposed method 87.65% accuracy was achieved.

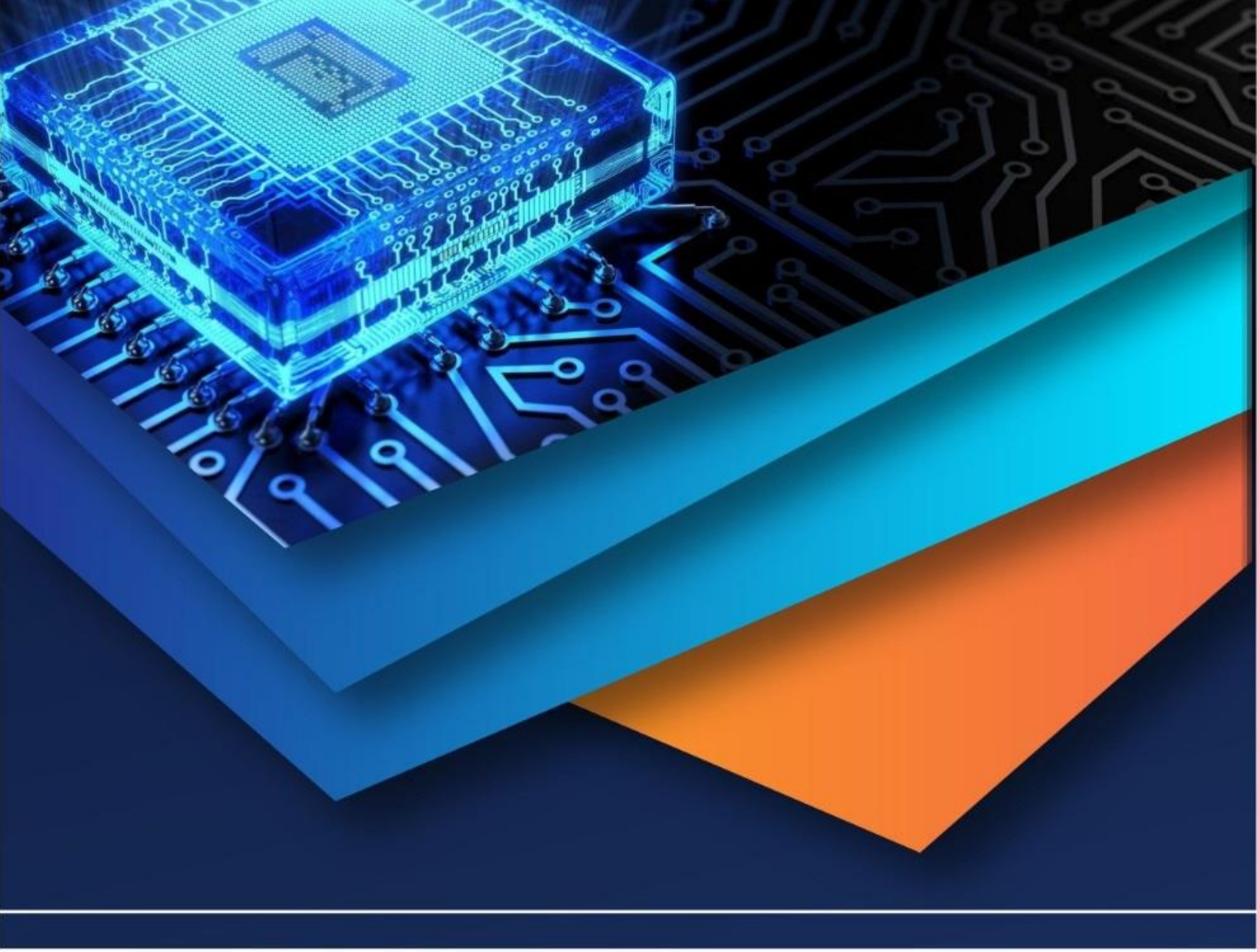
V. CONCLUSION

The proposed method calculates polarity scores and classified the COVID-19 tweets into Positive, Negative and Neutral category using TextBlob library. Multinomial Naive Bayes classifier was build to predict the new tweets into one of the 3 emotion category. Proposed method achieved an accuracy of 87.65%. From this research we can conclude that, during the covid-19 pandemic situation people are more neutral towards the situation, and this data can be used by the government and other organisation to make policies in the current situation and this analysis will help them to handle this type of situations in future. In future this analysis can be done using VADER (Valence Aware Dictionary and sEntiment Reasoner) Sentiment Analysis. Different machine learning classifiers can be build and compare the results.

REFERENCES

[1] E. Severeyn, S. Wong, H. Herrera, A. L. Cruz, J. Velásquez and M. Huerta, "Study of Basic Reproduction Number Projection of SARS-CoV-2 Epidemic in USA and Brazil," 2020 IEEE ANDESCON, Quito, 2020, pp. 1-6, doi: 10.1109/ANDESCON50619.2020.9272081.
[2] G. Subramaniam, R. Aswini, M. Ranjitha and P. K. Rajendran, "Survey on user emotion analysis using Twitter data," 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), Chennai, 2017, pp. 998-1001, doi: 10.1109/ICECDS.2017.8389587.
[3] 4. M. Rathi, A. Malik, D. Varshney, R. Sharma and S. Mendiratta, "Sentiment Analysis of Tweets Using Machine Learning Approach," *2018 Eleventh International Conference on Contemporary Computing (IC3)*, Noida, 2018, pp. 1-3, doi: 10.1109/IC3.2018.8530517.
[4] https://textblob.readthedocs.io/en/dev/
[5] A. P. Jain and P. Dandannavar, "Application of machine learning techniques to sentiment analysis," *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, Bangalore, 2016, pp. 628-632, doi: 10.1109/ICATCCT.2016.7912076.
[6] M. Sethi, S. Pandey, P. Trar and P. Soni, "Sentiment Identification in COVID-19 Specific Tweets," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2020, pp. 509-516, doi: 10.1109/ICESC48915.2020.9155674.

[7] Nithyashree Tand Nirmala M.B. "Analysis of the Data from the Twitter account using Machine Learning," 2020 5th International Conference on Communication and Electronics Systems (ICCES), COIMBATORE, India, 2020, pp. 989-993, doi: 10.1109/ICCES48766.2020.9137955.
[8] P. P. Surya and B. Subbulakshmi, "Sentimental Analysis using Naive Bayes Classifier," 2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN), Vellore, India, 2019, pp. 1-5, doi: 10.1109/ViTECoN.2019.8899618.
[9] https://ieee-dataport.org/open-access/coronavirus-covid-19-tweets-dataset
[10] S. Zahoor and R. Rohilla, "Twitter Sentiment Analysis Using Lexical or Rule Based Approach: A Case Study," 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2020, pp. 537-542, doi: 10.1109/ICRITO48877.2020.9197910.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY