iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 10 **Issue:** IV **Month of publication:** April 2022

**DOI:** https://doi.org/10.22214/ijraset.2022.41449

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 IV Apr 2022- Available at www.ijraset.com*

# Sentiment Analysis and Predictions of COVID 19 Tweets using Natural Language Processing

Prashant Chauhan[1], Amit Kumar[2], Pranjal Srivastava[3], Rahul Prajapati[4]
[1, 2, 3, 4]*Electronics and Communication Engineering, KIET Group of Institution, Ghaziabad, India*

***Abstract:*** ***Sentiment analysis is one of the major tasks of NLP. In this paper, we analyzed sentiment of covid-19 tweets using popular data science tools and divided this data into three categories- positive, negative, and neutral. we prepared this model to predict the category of the tweet using different supervised classifiers and examined their accuracy.***
***Keywords: NLP, Covid-19, Sentiment Analysis, TextBlob, KNN (K-Nearest Neighbor), RF (Random Forest), NB (Naïve Bayes) classifiers***

## I. INTRODUCTION

Sentiment analysis is an important area of research in NLP. In sentiment analysis, we analyse people's opinions, sentiments, belief, evaluations, attitudes, and emotions from written human language. Microblogging websites have a lot of information. People give a real-time opinion about various topics on these websites. The opinion can be helpful in different fields. Companies are using this information to build their strategy and manufacture their product. It can be used by policymakers or politicians. They can analyse public sentiments with respect to policies, public services, or political issues. It may be used to analyse the overall mood of public on a particular social issue. Sentiment analysis can help in prediction of an event.
This paper presents the result of sentiment analysis of covid-19 tweets and gives the accuracy of different Supervised ML classifiers that we used to predict the category of tweets. For the classification of tweets category, we used pre-existing sentiment classification library. We trained the model to predict the category of tweets using different classifiers and analysed the accuracy of the model with respect to different classifiers.

## II. METHODOLOGY

The methodology for this model consists of the following sections. These are-

*1)* Data collection
*2)* Pre-processing and feature extraction
*3)* Sentiment Analysis
*4)* Training and classification using different supervised ML classifier
*5)* Results

Figure 1 shows the flow chart of the methodology used in this model. It takes Row data, applied NLP technique for feature extraction. And trained with different ML classifier to predict the category of the given sentence.

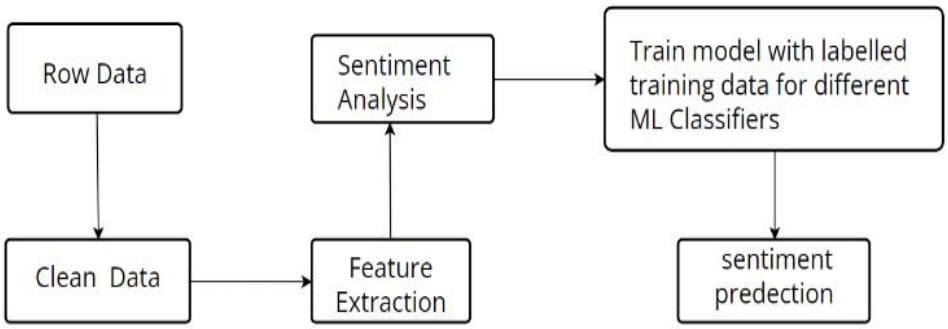



Fig. 1. Flow chart of the purposed methodology for the model

## III. DATA COLLECTION

In an effort to start working towards the main goal, finding proper data sources was the first step. There is a vast existing landscape of microblogging websites that could be used, but collecting data from different social media platforms was very time-consuming. Tweepy is a reliable, direct-from-source API for pulling data from Twitter. Same as Tweepy, for different microblogging platforms different APIs were needed to pair the social media sentiment about pandemic but in respect to time, it was determined to choose one platform that would provide us with a sufficient amount of data as well as a high ease-of-use for that data.

Kaggle was one for us. Kaggle has many data set for practicing data science projects. We used one of these data set which consists of more than 107k covid tweets and necessary information regarding this dataset. This dataset has total 610 sources from which we can observe top 10 sources of data in figure 2.

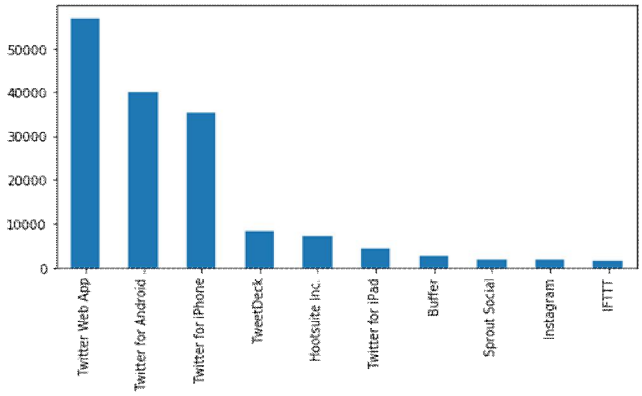



Fig. 2. Top 10 sources used for data collection

## IV. PRE-PROCESSING AND FEATURE EXTRACTION

After collecting the dataset for the model, the next step was to pre-process the data set and extract the features from the dataset. For this task, we used 'neatText' an NLP library, and python3.7. In this step, the dataset goes under the following process-

1) *Remove noise from Data: In this section, we removed unwanted data from the dataset. For example, '#Covid19' becomes 'Covid19'. In this section, we removed user names, URLs, hashed words, punctuation, special characters, etc. For papers with less than six authors: To change the default, adjust the template as follows.*
2) *Remove multiple-space:* After removing noise from data, extra space was generated. This can cause Vocabulary mismatch so the next target was to remove these extra spaces.
3) *Remove stop-words:* Stop-words are those words that generally do not contain important information about the tweets. For example, it can be a preposition like On, in, below, under, etc. It can be any article like a, an, the, etc. These words are used to make the sentence grammatically correct but these are found in abundance in a sentence so these stop-words should be removed before keyword extraction.
4) *Key-word extraction:* Keyword extraction is the technique to automatically extract important and pre-defined relevant words from the sentence. We used 'textBlob' an NLP library, for keyword extraction.
5) Sentiment classification: For the classification of sentences, we used 'textBlob' an NLP library and extracted the polarity of the sentences. According to the polarity of sentences, we categorized tweets into three categories which are positive, neutral, and negative. We can see the no. of tweets categorized in each category in figure 3.

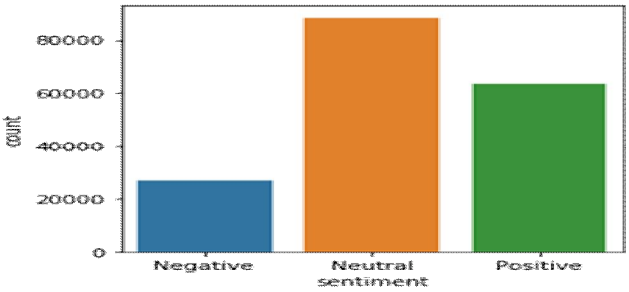



Fig. 3. Classification of sentiment

6) *Token Generation:* In the next step, we split the sentences into the words known as a token, and this process is known as tokenization. This helps in understanding the meaning and context of the sentences. we generated positive, negative, and neutral tokens.
7) *Word Cloud Formatio:* In the next step, we combined these tokens to generate a word cloud for the positive, negative, and neutral class sentiment. It helps in visualizing the most frequent words used in sentences.

## V. TRAINING AND CLASSIFICATION USING DIFFERENT SUPERVISED LEARNING ALGORITHMS

We used pyhton3, Jupyter notebook, and textBlob library for the NLP technique. The data set is processed using text blob and we got the polarity of tweets. We used this polarity to label the positive, negative, and neutral classifications. For the prediction of tweets, we used the labelled dataset, trained this dataset with different supervised ML classifiers, and analyzed the performance result of these classifiers on the model. Here, we used Random Forest classifier, Naïve bayes classifier, KNN.

## VI. RESULT DISCUSSION

We used 179k tweets to classify the sentiment of the tweet. For prediction purposes, we used 80k classified tweets to train the model and then tested it using three supervised ML classifiers. These are the RF, NB, and KNN classifier. The system is tested for this four parameters-Accuracy, Precision, Recall, F score.

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ negative + False\ Positive + False\ Negative} \quad (1)$$

(1

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (2)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (3)$$

$$F\text{-}Score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (4)$$

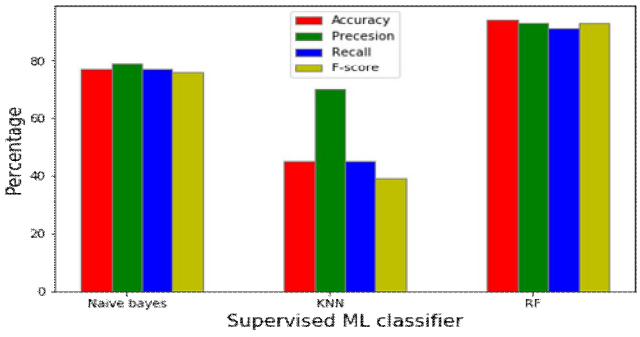



Fig. 4. Performance with respect to different ML classifier

The figure 4 shows the performance of the different supervised classifiers on the data set. From the figure, it can be observed that the RF classifier is most accurate for the classification of sentences and KNN is the least accurate. Whereas the performance of the NB classifier lies between RF and KNN classifier.

## VII. CONCLUSION AND FUTURE WORK

This model applies NLP techniques and extracts features from tweets. These features are used for the classification of tweets in positive, neutral, and negative sentiment. Then it compares the performance of RF, KNN, and NB classifiers for the prediction of the class of tweets. We can observe the poor performance of KNN. In future work, we plan to decrease the computational cost and connect our model with deep learning so that its performance can get increased.

## REFERENCES

[1] In Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020. 2020.Bose, Rajesh, P. Aithal, and Sandip Roy. "Sentiment analysis on the basis of tweeter comments of application of drugs by customary language toolkit and textblob opinions of distinct countries." Int. J 8 (2020).

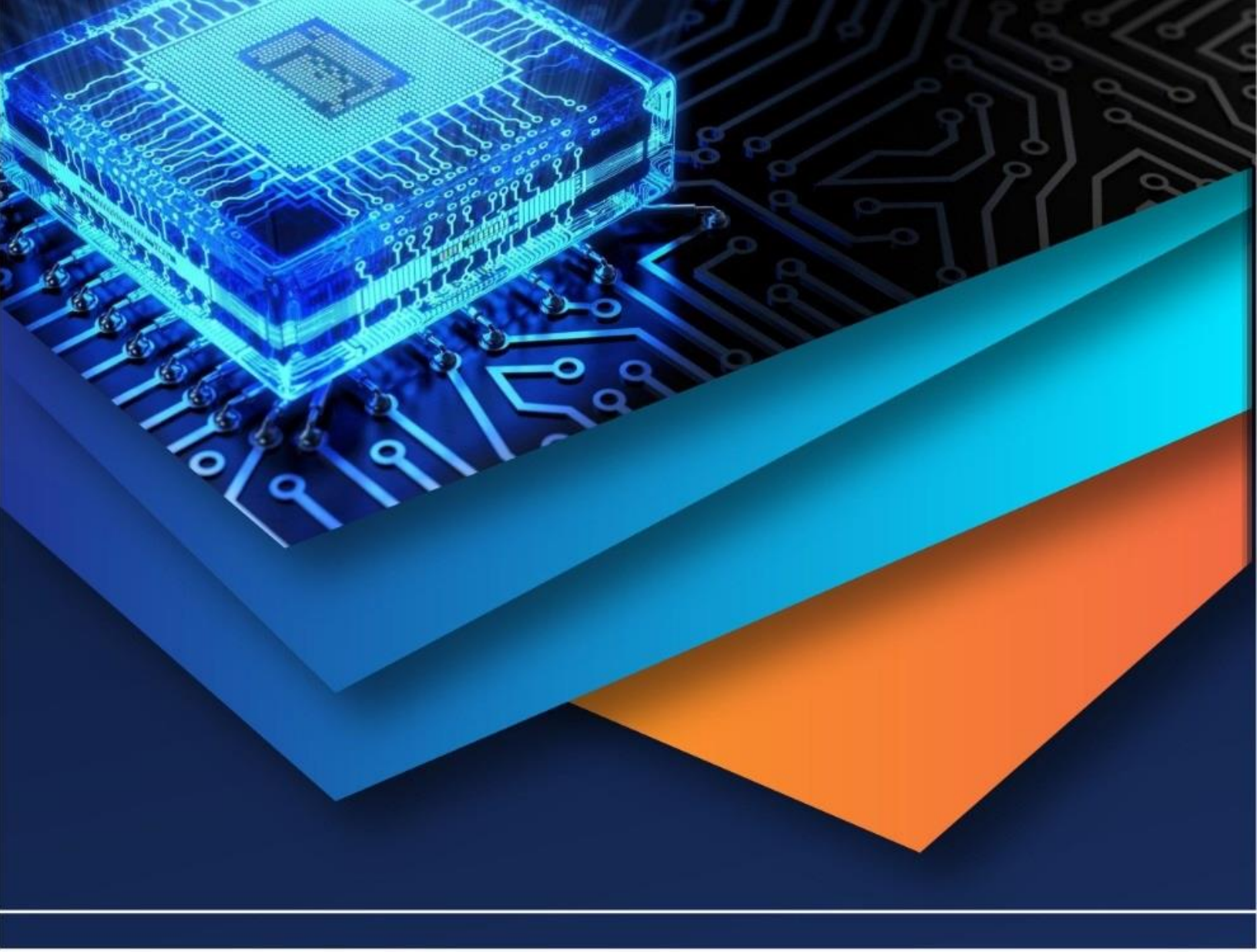
[2] Xu, Jin, Yubo Tao, and Hai Lin. "Semantic word cloud generation based on word embeddings." In 2016 IEEE Pacific Visualization Symposium (PacificVis), pp. 239-243. IEEE, 2016.

[3] Serrano, Juan Carlos Medina, Orestis Papakyriakopoulos, and Simon Hegelich. "NLP-based feature extraction for the detection of COVID-19 misinformation videos on YouTube."

[4] Untawale, Tejaswini M., and G. Choudhari. "Implementation of sentiment classification of movie reviews by supervised machine learning approaches." In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), pp. 1197-1200. IEEE, 2019.

[5] Untawale, Tejaswini M., and G. Choudhari. "Implementation of sentiment classification of movie reviews by supervised machine learning approaches." In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), pp. 1197-1200. IEEE, 2019.

[6] Alemzadeh, H. and Devarakonda, M., 2017, February. An NLP-based cognitive system for disease status identification in electronic health records. In 2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI) (pp. 89-92). IEEE.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY