IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 8, April 2024

# Prevalent Disease Tweet Classification using Sentimental Analysis

**Kevin Kurien Joseph[1], Rahul S[2], Siddhaarthan V[3], Indumathy[4]**
Department of Computer Science and Engineering[1,2,3,4]
SRM Institute of Science and Technology, Vadapalani, Chennai, India

**Abstract***: Interest in people's opinions about tweets has lately grown as a result of opinionated data that can be obtained on blogs and social media platforms. With everyone having access to the internet and the ability to tweet whatever they want, Twitter is one of the most widely used information sources in the world today. Therefore, there is a higher likelihood that people would be misled. It would be very helpful to come to a consensus on what the general attitude of the populace is, particularly during times of panic like the pandemic, in order to better grasp how ready the populace is to face such a potentially disastrous crisis. To handle these kinds of challenging problems, machine learning methods are typically applied. Analysing this kind of data manually requires more time. It is challenging to categorise opinions according to their polarity. This project was created specifically to analyse the moods of Twitter users using ML techniques and NLP tools to pinpoint the cause of COVID-19's reappearance. The perspectives of different people can diverge. After the system can classify the text's emotions, we train it using historical data. Machine learning methods are crucial to this project's ability to classify data. The Random Forest Classifier has also been chosen as the framework's preferred ML approach, with an accuracy rate of 88%. We assessed the framework using a variety of metrics, including F1-Score, precision, accuracy, and recall*

**Keywords:** precision, accuracy

## I. INTRODUCTION

A variety of ideas and feelings were sparked by the Coronavirus's widespread distribution. Due to the sheer nature of the COVID-19 pandemic, confusion and panic spread quickly through society. People from many countries responded on social networking networks in different ways. The COVID-19 epidemic has helped to highlighting urban dwellers' vulnerabilities and constitutes a significant public health risk. When there was a pandemic, people's emotions varied, which led to mental problems including fear, worry, and many other horrifying symptoms. Twitter posts including the phrases "updates about confirmed cases," "COVID-19-related death," "early signs of the outbreak," "economic impact," and "preventive measures" reveal feelings of dread and fear. Additionally, Public thoughts about COVID-19 news posted on microblogging platforms have the power to propagate different points of view. Natural language processing technologies are used., we describe an emotional analysis of COVID-19 in this work. By examining news articles and social media data, we hope to assess how the general public is feeling about COVID-19. We'll examine the feelings people have about the pandemic and pinpoint the key issues and themes that are influencing perception. User-generated content on social media is significant since it can serve as a significant information source during times of crisis. Social media and microblogging platforms like Facebook and Twitter were massively accepted by people. Since the majority of the data were from the early 2020s to the late 2021s, selecting a dataset at first presented us with some difficulties. However, we only selected our dataset from the more than 2500 English-language tweets that were sent on Covid-19 during October and December of 2022. From the word corpus, we identified the most widely used words. Additionally, Using the sentiment scores of our preprocessed tweets, we classified our dataset and assigned sentiment scores to the tweets based on the polarity of their sentiment. Finally, we trained our model using those tweets and their sentiment score.

IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.53

Volume 4, Issue 8, April 2024

## II. MOTIVATION

We wanted to broaden our focus in order to better understand how the world felt at the time about Covid-19 in light of the unexpected increase of instances that is currently occurring in China. to use Twitter to gather consensus on how the world felt about the possibility of its revival. By reaching a consensus, we hope to convey to the user the general opinion held by those who have posted the tweet or other related tweets about the widespread illness. to comprehend how people throughout the world feel about another possible global shutdown. The COVID-19 pandemic has also generated a great deal of false information and fake news, which has left people perplexed and afraid.Sentiment analysis can assist in identifying such inaccurate and misleading information and in eradicating it by supplying the public with reliable and accurate information

## III. RELATED WORK

| | Literature Survey | | |
|---|---|---|---|
| | Reference | datasets | method |
| 1 | Birgül Öneç, Nilüfer Yurtay, Tuba Karagül Yıldız | from the Düzce University Research and Application Hospital | Four different machine learning techniques are used to categorise the 12 forms of anaemia that are most frequently found in the province of Düzce: naive Bayes, ANNs, SVMs, and ensemble decision trees |
| 2 | Jinglan Zhang, Ye Duan, Mohammed A. Fadhel, Omran Al-Shamma, Laith Alzubaidi | ImageNet dataset | The classification of sickle cell illness has performed much better because to the same domain transfer learning. |
| 3 | ManisshJaiswal, Anima Srivastava. Tanveer J, Sidiqui | data gathered from regional pathology facilities | The experimental outcome on a sample dataset reveals that, when compared to C4.5 and Random forest, the Naive- Bayes classification algorithm performs better in terms of accuracy. |
| 4 | .RajanVohra, AbirHussain. Anil Kumar Dudyala, JankiSharanPahariya | Mendeley Data | Using the same data sets, a novel hybrid model based on Deep Learning, Genetic Algorithms, and Convolutional Neural Networks (CNN) may be created for classifying Iron Deficiency Anaemia. |
| 5 | Bhimsen, Adarsh Ganesh AnupamaBhan, Shubhra Dixit, AyushGoyal | Basic blood reports | The photos are retrieved for their geometrical, statistical, and textural properties.We use the machine learning classifiers of support vector machine, logistic regression, and random forest.. These algorithms' comparisons are described in this study. |
| 6 | Shilpa A. Sanap,MeghanaNagori&VivekKshirsagar | created using complete blood count test results from different hospitals | Based on CBC reports and anaemia severity, a decision tree for anaemia classification is created that provides the best anaemia classification feasible. We have seen that the C4.5 algorithm performs best and is most accurate. |
| 7 | El-AwadyAttia , ArifIqbal Umar, Dr. Syed Hamad Shirazi, Zakir Khan, Asfadyar | ShaukatKhanum Hospital and Research Centre helped create a | The overall findings show that the proposed model had recall F1-Score and specificity of 98.95%, 98.12%, and 98.12% for training, validation, and test accuracy, |

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 8, April 2024

| | | | |
|---|---|---|---|
| | Khan, Muhamad Assam, Abdullah Mohamed , Muhammad Shahzad | healthy RBC dataset. | respectively, of 91.37%, 88.85%, and 86.06%. |
| 8 | Sahar J. Mohammed; Amjad A. Ahmed; Mohammed Sami MOHAMMED, Arshed A. Ahmad; | 539 participants' data, with 10 relevant attributes for each, were gathered for this study. | To create a curate anaemia prediction system, three based rule classification techniques are used: ZeroR, OneR, and PART to collect pertinent anaemia datasets connected with "If" and "Then" procedures. In comparison to ZeroR and OneR, PART's accuracy was 85% greater. |
| 9 | SamikshaSoni; HardikThakkar; Bikesh Kumar Singh | A recently created database with 67 features and 23 illness cases is used for the investigation. | With a 10-fold data division protocol, the suggested system displays the greatest classification accuracy of 95.5%. Precision, sensitivity, specificity, negative projected value, and ROC curve are additional performance metrics utilised for evaluation. |
| 10 | PoojaTukaramDalvi; NagarajVernekar | the National Health and Nutrition Examination Survey (NHANES) datasets for 2007–2008 and 2009–2010. The dataset contains information on demographic, health, and laboratory measures for a nationally representative sample of the non institutionalized civilian US population | Among the several classifiers, the Artificial Neural Network performs the best, while K-Nearest Neighbour performs the worst. The classifier combination of Decision Tree and K-Nearest Neighbour, however, produces a significantly greater accuracy than the Artificial Neural Network when applied to a stacking ensemble. |

Table 1

Contribution:

By offering a fresh sentiment classification model on howtocategorise people's emotions those are frequently expressed on microblogging websites, this paper adds to the body of information already available regarding COVID-19 tweet classification. Hereare our main contributions.

1) In terms of data preprocessing, we created a novel method called The PorterStemmer that uses phoentics rather than more traditional techniques like lemmatization to analyse and clean the data. Modern ML methods for tweet classification are used in conjunction with the newly introduced preprocessing methodology. The Random Forest Classifier beats the manual tweet classification challenge, as shown by our architecture.

2) In terms of performance, We have employed a range of assessment metrics, such as accuracy, precision, recall, and F1 score, to validate the rationale behind our model design.. We show how to use a methodology to compare the top methods using several evaluation criteria. Our investigation of several ML methods and how well they function in relation to COVID-19 sentiment categorization advances knowledge of practical AI applications. We tweaked our hyperparameters to the nth degree before settling on the best hyperparameters, which gave our RandomForestClassifier the best accuracy of 88.03%.

Our strategy of cleaning the tweets and utilising a phonetics-based stemming algorithm can significantly increase the scalability of text processing for sentiment analysis because just tweaking the hyperparameters will hinder us from scaling up our model successfully.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 8, April 2024

## IV. METHODOLOGY

Our approach is distinctive because it utilises a hybrid architecture that blends supervised machine learning methods for categorising tweets with NLP-based methods for sentiment analysis and tagging. In order to identify the best supervised ML method that generates high accuracy and recall, we employed and tested a range of supervised ML algorithms. Cross-

validation grid-search was used to change the parameter values in order to optimise the performance of the ML approaches. The classification method we have developed is then reviewed.

**Work flow diagram**

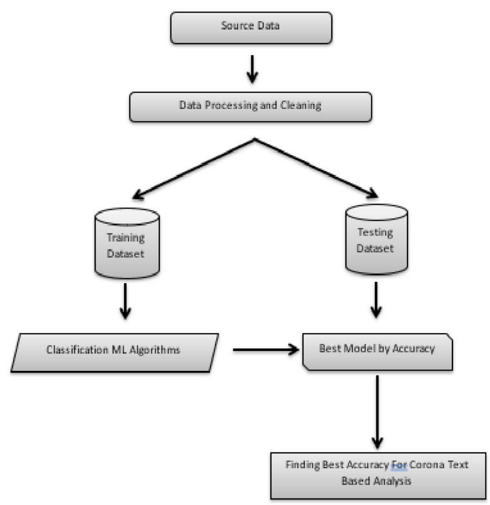


Fig 1Architecture Diagram

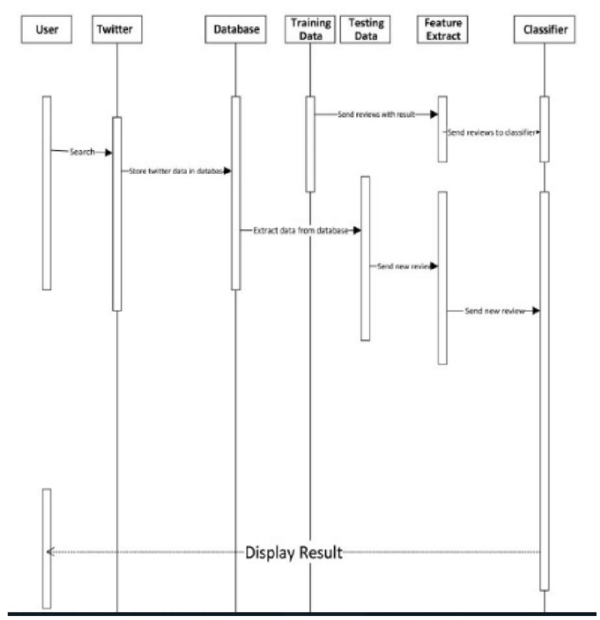


Fig.2SequenceDiagram

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 8, April 2024

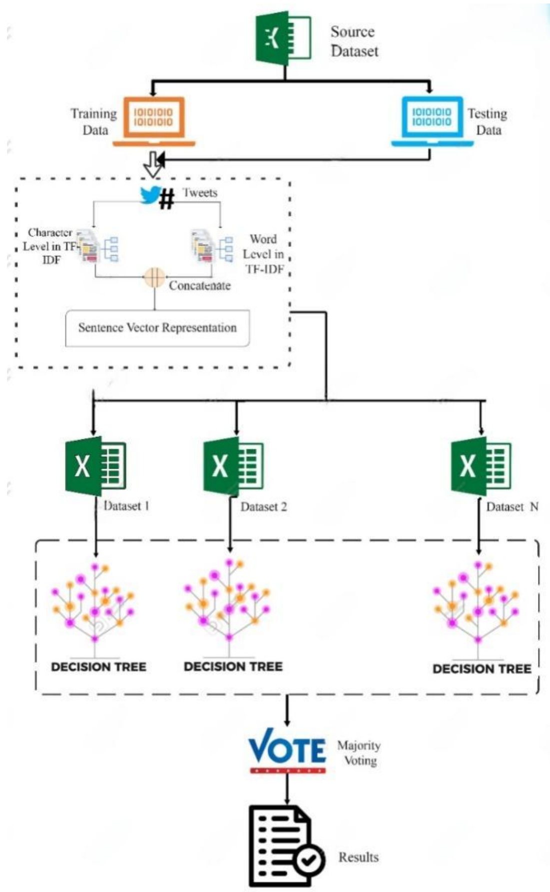


Fig 3

**A. Model**

**Data Pre-processing**

1) With the help of the Python-based NLP toolkit (NLTK), we have eliminated stop words like "over," "under," "again," "further," "then," "once," and "there." Using regular expressions (RegEx), we also eliminated URLs, punctuation (*, $, and! ), user mentions (@), and hashtags (#).

2) Case Folding: A term's polarity (positive, negative, and neutral) can be significantly influenced by the uppercase and lowercase variations of that word. Depending on the nature of the corpus and its intended use, some words in the text may be changed from uppercase to lowercase or vice versa. Our dataset comprises of short-worded tweets that typically use lowercase and follow a standard communication style. In order to do case-folding, we reduced the number of capital words in all tweets.

3)Tokenization: We broke down each tweet message into tokens (words) so that machine learning (ML) algorithms could grasp the meaning of each word.

4) Stemming, also referred to as the base or root form, is the process of reducing a word to its stem. Stemming seeks to normalise words such that related terms sound similar. (like "run", "runs", and "running") are viewed as one and the same. This can help text analysis tasks like sentiment analysis, information retrieval, and text classification perform better.

**Data Visualization:**

Data visualisation is a vital skill in applied statistics and machine learning.. Statistics actually puts a lot of emphasis on numerical estimations and data descriptions. Data visualisation offers a crucial set of tools for gaining a qualitative insight.. This can aid in understanding the data distribution of the dataset. Visualisations can be used to express and convey crucial linkages through plots and charts that are more immediately and pertinent to stakeholders rather than using measurements of association or relevance. Even checking for training dataset data that are unrelated to the model can be done using it.

**Classifiers:**

Since we automatically classified the tweets based on the sentiments using supervised ML approaches, these techniques are known as classifiers.Each algorithm needs to be assessed uniformly on the same data is essential for conducting a fair comparison of machine learning algorithms, The positive, negative, and neutral attitudes are the fundamental classes for their subclasses (i.e., extremely positive, positive, neutral, negative, and highly negative). As tweets make up the majority of the used data, the limited-class sentiment extraction was used. Through a process of testing various ML modules that have been prominently been used for Multiclass Classification (eg. Logistic Regression, Decision Tree, LSTM etcWe identified three machine learning (ML) algorithms that outperformed the others. These were:

- Logistic Regression
- Catboost Classifier
- Random Forest Classifier

They provided accuracy scores of 84.3%, 85%, and 88.03%, respectively.

## V. EXPERIMENTAL RESULTS

Despite what its name might imply, the logistic regression statistical classification method produces the classification probability of the desired output (the dependent variable).. Using the collected features, the logistic regression model is trained using the labelled dataset of tweets. The model gains the ability to modify the weights of the features to reduce the discrepancy between its predictions and the actual labels in the training set. The output of the linear model is passed via the sigmoid function after the logistic regression model has been trained, which converts the output to a probability value between 0 and 1. The sigmoid function has the following formula:

$$1 / (1 + \exp(-z)),$$

Finally, the sentiment of a given tweet is predicted using the probability value returned by the sigmoid function.

```
from sklearn.linear_model import LogisticRegression
lr = LogisticRegression()
lr.fit(X_train,y_train)

LogisticRegression()
In a Jupyter environment, please rerun this cell to show the HTML representation or trust the noteboo
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.o

lr=LogisticRegression()
param_grid = {
    'C':[0.001, 0.01, 0.1, 1, 10, 100, 1000]

    }
CV_lr = GridSearchCV(estimator=lr, param_grid=param_grid, verbose=3)
CV_lr.fit(X_train, y_train)

GridSearchCV(estimator=LogisticRegression(),
             param_grid={'C': [0.001, 0.01, 0.1, 1, 10, 100, 1000]}, verbose=3)
In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

CV_lr.best_params_

{'C': 10}

lr=LogisticRegression(C=10)
lr.fit(X_train,y_train)

LogisticRegression(C=10)
In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

predict = lr.predict(X_test)
```

Fig 4

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 8, April 2024

```
from sklearn.metrics import classification_report
print('Classification report of Logistic Regression\n\n',classification_report(y_test,predict))
```

```
Classification report of Logistic Regression

                   precision    recall  f1-score   support

Extremely Negative      0.83      0.65      0.73        77
Extremely Positive      0.88      0.75      0.81        77
          Negative      0.78      0.92      0.85        77
           Neutral      0.85      0.92      0.88        77
          Positive      0.92      1.00      0.96        77

          accuracy                          0.85       385
         macro avg      0.85      0.85      0.84       385
      weighted avg      0.85      0.85      0.84       385
```

Fig 5

```
cm1=confusion_matrix(y_test,predict)
sns.heatmap(cm1,annot=True,cmap='coolwarm')
plt.title('Confusion Matrix-Logistic Regression')
```

```
]: Text(0.5, 1.0, 'Confusion Matrix-Logistic Regression')
```

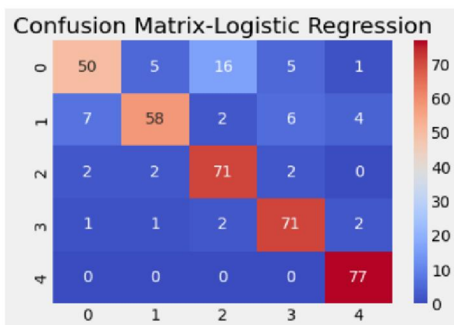


Fig 6

**Catboost**

CatBoost is a gradient boosting method that can be used for sentiment analysis and other binary or multi-class classification applications. Through the iterative construction of a number of decision trees that gauge the emotion of the tweet based on the input features, CatBoost learns to minimise the log loss or cross-entropy loss function during training. The input features that were most crucial for predicting the sentiment of the tweets can be revealed using CatBoost. This can be helpful for figuring out which elements of the tweets have the most effects on sentiment. The handling of categorical features, the handling of missing values, and the provision of feature importance analysis are only a few of CatBoost's benefits for sentiment research. CatBoost also has quick training times and can handle huge datasets.

```
#Catboost
from catboost import CatBoostClassifier
```

```
from catboost import CatBoostClassifier
CB = CatBoostClassifier()
CB.fit(X_train,y_train)
```

```
Learning rate set to 0.084304
0:      learn: 1.5888496        total: 327ms    remaining: 5m 26s
1:      learn: 1.5722639        total: 465ms    remaining: 3m 51s
2:      learn: 1.5571412        total: 602ms    remaining: 3m 20s
3:      learn: 1.5432026        total: 738ms    remaining: 3m 3s
4:      learn: 1.5316613        total: 873ms    remaining: 2m 53s
5:      learn: 1.5213331        total: 1.01s    remaining: 2m 47s
```

```
pred2=CB.predict(X_test)
```

```
pred2
```

Fig 7

IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.53

Volume 4, Issue 8, April 2024

```
print('Classification report of Cat boost Classifier\n\n',classification_report(y_test,pred2))
```

```
Classification report of Cat boost Classifier

                   precision    recall  f1-score   support

Extremely Negative      0.76      0.71      0.74        77
Extremely Positive      0.86      0.81      0.83        77
          Negative      0.86      0.77      0.81        77
           Neutral      0.81      0.96      0.88        77
          Positive      0.95      1.00      0.97        77

          accuracy                          0.85       385
         macro avg      0.85      0.85      0.85       385
      weighted avg      0.85      0.85      0.85       385
```

```
cm2=confusion_matrix(y_test,pred2)
sns.heatmap(cm2,annot=True,cmap='coolwarm')
plt.title('Confusion Matrix-Catboost')
```

```
Text(0.5, 1.0, 'Confusion Matrix-Catboost')
```

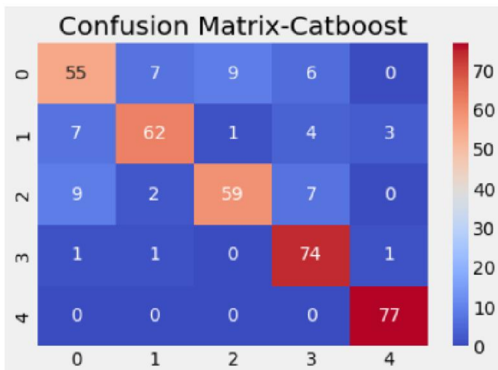


Fig8

**Random Forest Classifier**

Data scientists frequently employ the algorithm Random Forest. A popular supervised machine learning method for classification and regression issues is random forest.

The Random Forest combines different Decision Trees to provide an output that was chosen by all trees. The Random Forest method does not overfit and is quick, scalable, noise-reduced, and accurate.

Important Random Forest Features

• Diversity: Not all characteristics, elements, or features are taken into account when building a particular tree because each treeis unique

• Immune to the dimensionality curse: The feature space is reduced since each tree only considers a portion of the data.

• Parallelization: Each tree is individually generated using different data and attributes. This means that building random forests can utilise the CPU to its fullest potential.

• Stability: There is stability since the result is based on a majority vote or an average.

The Random Forest Classifier runs new input text data through each of the forest's decision trees. Each decision tree offers a forecast based on the tenor of the input text data. In order to arrive at the final forecast, all of the decision trees in the forest's probabilities are averaged, or their projections are combined.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 8, April 2024

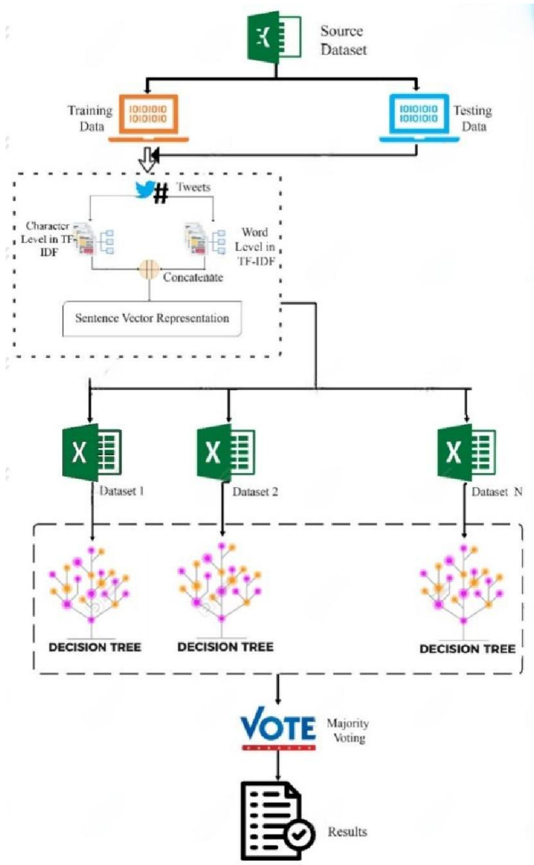


Fig 9

With an accuracy rating of 88.03%, Random Forest Classifier was picked as the algorithm to use in building our model

```
rfc=RandomForestClassifier(random_state=47,n_estimators=147,max_depth=None,min_samples_leaf=1,min_samples_split=5)
rfc.fit(X_train,y_train)

RandomForestClassifier(min_samples_split=5, n_estimators=147, random_state=47)
In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

rfc_pred=rfc.predict(X_test)

from sklearn.metrics import accuracy_score
forest=accuracy_score(y_test,rfc_pred)*100
print('Accuracy of Random Forest Classifier',accuracy_score(y_test,rfc_pred)*100)#random_state=32

Accuracy of Random Forest Classifier 88.05194805194805

from sklearn.metrics import classification_report
print('Classification report of Random Forest Classifier\n\n',classification_report(y_test,rfc_pred))

Classification report of Random Forest Classifier

                   precision    recall  f1-score   support

Extremely Negative      0.77      0.74      0.75        77
Extremely Positive      0.88      0.82      0.85        77
          Negative      0.95      0.91      0.93        77
           Neutral      0.82      0.94      0.87        77
          Positive      1.00      1.00      1.00        77

          accuracy                          0.88       385
         macro avg      0.88      0.88      0.88       385
      weighted avg      0.88      0.88      0.88       385
```

Fig 10

```
cm3=confusion_matrix(y_test,rfc_pred)
sns.heatmap(cm3,annot=True,cmap='coolwarm')
plt.title('Confusion Matrix-RandomForest Classifier')
```

```
Text(0.5, 1.0, 'Confusion Matrix-RandomForest Classifier')
```

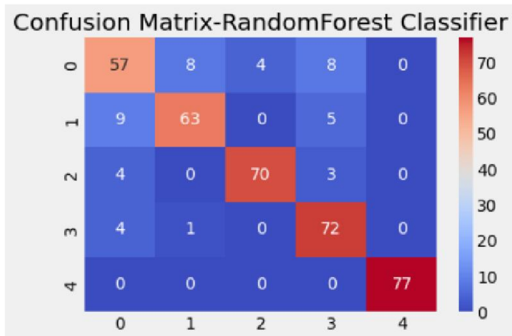


Fig 11

## VI. PERFORMANCE EVALUATION

### Performance Metrics

With the help of the Using the performance measures recall, accuracy, and precision as well as F1-score classification, we evaluated the efficiency of the applied ML techniques. May TP represent the total number of samples that were positively recognised or the total number of true positives. True Negatives is the initials TN., or the quantity of incorrectly categorised negative samples. "False positive" (FP) samples are those that were mistakenly identified as positive despite being negative. "False negatives" (FN) are the number of positive samples that were mistakenly categorised as negatives.

Precision measures overall efficacy. The easiest to understand is the accuracy performance statistic, which is simply the proportion of properly predicted observations to all observations. Given that we have a high level of accuracy, our model is the best. The datasets' false positive and false negative rate values must be roughly comparable for accuracy to be a good indicator.

The accuracy formula is (TP+TN)/(TP+FP+FN+TN).

The accuracy of the results generated is evaluated. In terms of positive observations, precision is the ratio of accurately predicted observations to all expected positive observations. Low false positive rates are inversely associated with precision.

Precision is defined as TP/(TP+FP).

The quantity of tweets remembered that were categorised correctly. Recall (Sensitivity) - Recall measures the proportion of positive observations among all actual class observations that were accurately predicted.

Recall is TP / (TP + FN).

A weighted average of recall and precision scores is the F1-Score. It averages recollection and precision. As a result, both false positives and false negatives are considered while calculating this score. F1 is often more advantageous than accuracy, especially if you have an uneven class distribution, although it is not as intuitively simple to understand as accuracy. When false positive and false negative costs are about similar, accuracy performs best. The ideal approach is to consider both Precision and Recall.

F1 score is calculated as (2(precision recall))/(precision + recall).

F-Measure is defined as 2TP divided by (2TP plus FP plus FN).

Hyperparameter Tuning

Hyperparameters, is the crucial part of every ML technique, are frequently in charge of significant performance improvements. The hyperparameters can be modified using crossvalidation grid search, which uses the to get the best settings for the hyperparameters, utilise either the training dataset or the whole dataset. In the cross validation, we select the set of hyperparameter values that produce the maximum level of accuracy by using accuracy as the performance metric. The relevant ML algorithm is then implemented using these hyperparameter values.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 8, April 2024

Hyper Parameter Tuning For Each Of The Ml Algorithms

| Model | Type | Tuned Parameters | Description |
|---|---|---|---|
| Random Forest | Classifier | {'max_depth':None,'min_samples_leaf:1,'min_samples_split:5,'n_estimators: 147} | max_depth is the layers of each tree,min_sample_leaf is the mininum number of features in a node for it to be Min_sample_split is the bare minimal set of features required for a node to be split into the leaf node., The amount of decision trees in the Random forest is indicated by the variable n_estimators. |
| Logistic Regression | Classifier | {'C': 10} | C is a regularlization parameter it is used to stop Overfitting and underfitting |
| Catboost | Classifier | {'max_depth':5, 'n_estimators': 300} | The maximum depth of a decision tree is a hyperparameter that can be tuned by the user. It is controlled by the "max_depth" parameter and establishes the most levels a tree may have. The maximum number of trees that can be created during training is determined by the "n_estimators" option. |

Table 2

## VII. RESULT &CONCLUSION

This section discusses the overall experimental results thus obtained after comparing the machine learning classification models.

According to the Classification Accuracy figure the Accuracy of random forest is the highest which is 88.32%. The Accuracy of Catboost is the second highest which is 85.73%. The accuracy of Logistic Regression is slightly less than Catboost classifier ie 85.54

```
Accuracy['Algorithm']=Accuracy.index
Accuracy
```

| | Accuracy | Algorithm |
|---|---|---|
| **Logistic Regression** | 84.935065 | Logistic Regression |
| **Linear Regression** | 58.961039 | Linear Regression |
| **Random Forest** | 88.051948 | Random Forest |
| **Decision Tree** | 74.805195 | Decision Tree |
| **Catboost** | 84.935065 | Catboost |

The comparison of the accuracy score of each models are represented above. We can observe from the above table that the accuracy score of The Random Forest Model is the highest.

While through additional research the LSTM model showed comparable results to the Random Forest Classifier. Even though LSTM is more commonly used to handle NLP based problems, it is more prone to over fitting and extremely computationally expensive to train. Hence Random Forest Classifier has been chosen as the algorithm to be used to train our model. However, in the future if need to deal with a much larger dataset LSTM can be preferred since because they can learn long-term dependencies between inputs and outputs, which is often necessary in complex problems.

The model uses phonetics based stemming process, since lemmatization takes too long to train and the accuracy drop of isn't that significant. However in case of a smaller training dataset lemmatization could be used.

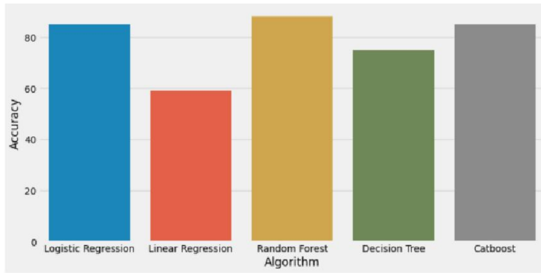


Fig 12

Data preparation and processing, missing value analysis, exploratory analysis, and model construction and evaluation came first in the analytical process. Random forest was shown to have the best accuracy on a public test set of higher accuracy score algorithms, with an accuracy of 88.32%. In order to reach a consensus about the overall sentiment embodied by all the tweets, it is employed in the application that can assist in identifying the sentiment hidden in the text of a tweet or a collection of tweets

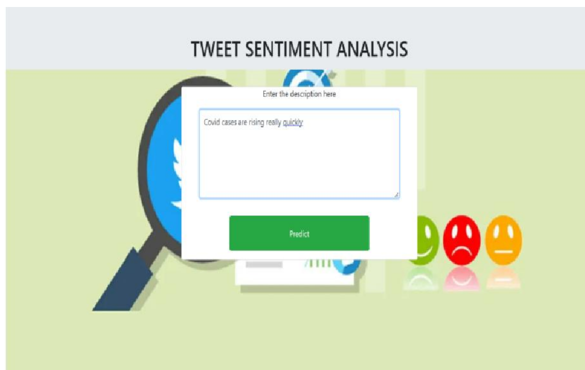


Fig.13

The model uses phonetics based stemming process, since lemmatization takes too long to train and the accuracy drop of isn't that significant. However in case of a smaller training dataset lemmatization could be used

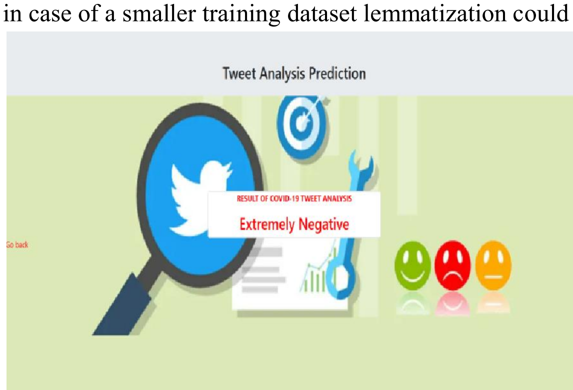


Fig. 14

## REFERENCES

[1]. NilufaYeasmin, NosinIbnaMahbub, MrinalKantiBaowaly, Bikash Chandra Singh, ZulfikarAlom, ZeyarAung and Mohammad Abdul Azim, "Analysis and Prediction of User Sentiment on COVID-19 Pandemic Using Tweets", Big Data Cogn. Comput. 2022, 6, vol 65

[2]. Arunava Kumar Chakraborty(RCC Institute of Information Technology,) , Sourav Das(AbulKalam Azad University of Technology), Anup Kumar Kolya(RCC Institute of Information Technology), "Sentiment Analysis of

Covid-19 Tweets using Evolutionary Classification-Based LSTM Model", 2022 15th IEEE International Conference on Recent Advances.

[3].SotiriaVernikou, AthanasiosLyras, Andreas Kanavos, "Multiclass sentiment analysis on COVID-19-related tweets using deep learning models", Published on 6th august 2022 on the springer- verlag London

[4].wethaSreeCheeti, Yanyan Li and Ahmad Hadaegh(California State University-San Marcos, San Marcos, California, USA), "Twitter based sentiment analysis of impact of covid-19 on education", International Journal of Artificial Intelligence and Applications (IJAIA), Vol.12, No.3, May 2021

[5]. DharmaiahDevarapalli, MedapatiSwapna Sri, PallemKavya Sri, PadmanabhuniCharishma, and PenmetsaVenu Naga Mounika, "Sentiment Analysis of COVID-19 Tweets Using Classification Algorithms", IEEE Journal of Biomedical and Health Informatics ( Volume: 24, Issue: 12, December 2020)

[6]. Immanuel, R. R., &Sangeetha, S. (2023). Identifying Different Emotions Of Human Using Eeg Signals Using Deep Learning Techniques. Journal Of Theoretical And Applied Information Technology, 101(18).

[7].Mohammad Abu Kausar ,ArockiasamySoosaimanickam , Mohammad Nasar(Department of Computing & Informatics, Mazoon College, Sultanate of Oman), "Public Sentiment Analysis on Twitter Data during COVID-19 Outbreak", International Journal of Advanced Computer Science and Applications, Vol. 12, No. 2, 2021

[8].Christina Knight, Chiara Biondi,Elyse Cornwall, "Twitter Sentiment Analysis: Global Attitudes Towards COVID-19 Policies", IEEE 4th international conference on geo science, networking and applications. Vol 12 2022

[9].TahaniSoudAlharbi, FethiFkih, "Building and Testing Fine-Grained Dataset of COVID-19 Tweets for Worry Prediction", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 13, No. 8, 2022

[10]. Rohitash ChandraID1☯*, Aswin Krishna2, "COVID-19 sentiment analysis via deep learning during the rise of novel cases", Chandra R, Krishna A (2021) COVID-19 sentiment analysis via deep learning during the rise of novel cases. PLoS ONE 16(8): e0255615.

[11].Rubul Kumar Bania, "COVID-19 Public Tweets Sentiment Analysis using TF-IDF and Inductive Learning Models", INFOCOMP, v. 19, no. 2, p. 23-41, December, 2020

[12]. Immanuel, R. R., &Sangeetha, S. K. B. (2023). Implementation of an Automatic EEG Feature Extraction with Gated Recurrent Neural Network for Emotion Recognition. In Computer Vision and Machine Intelligence Paradigms for SDGs: Select Proceedings of ICRTAC-CVMIP 2021 (pp. 133-150). Singapore: Springer Nature Singapore.

[13]. Muhammad Ali Shaikh, "Covid-19 Twitter Sentiment Analysis Using Machine Learning", 2022 University of Victoria

[14]. AdeolaAyandeyi, "SENTIMENT ANALYSIS OF TWITTER DATA TO ANALYZE THE EFFECT OF COVID-19", Concordia University of Edmonton 2021

[15]. Khairiyah Mohamed Ridhwan, Carol Anne Hargreaves, “Leveraging Twitter data to understand public sentiment for the COVID‑19 outbreak in Singapore”, International Journal of Information Management Data Insights 1 (2021) 100021

[16]. Jim Samuel , G. G. Md. Nawaz Ali , Md. MokhlesurRahman , EkEsawi and Yana Samuel, "COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification", Information 2020, 11, 314

[17]. RamadhanRenaldy;Muljono;GuruhFajarShidik;CaturSupriyanto;M.A. Soeleman; AhmadZainulFanani; Aditya DwiPratama, "Hybrid Method Based on Stacking for Sentiment Analysis of Indonesian Tweet Responding to COVID-19 Pandemic", 2021 International Seminar on Application for Technology of Information and Communication (iSemantic)