International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 9, pp: 694-701 www.ijaem.net ISSN: 2395-5252

Data Science tools for Exploratory Data Analysis,Time Series Analysis on the Covid19 Datasetand Coronavirus Tweets NLP Text Classification.

Apoorva R.
Master's student, Computer Science and Engineering,Centre for PG Studies, Mysore.

Date of Submission: 15-11-2020 Date of Acceptance: 30-11-2020

ABSTRACT: Exploratory data analysis is the process of analysing data to make sense of it and come up with a summary that describes the data using different methods like data manipulation and visualization. In other words, EDA is a way of describing and understanding the data beyond the values in rows and columns in the tables.Data Visualization is a graphical representation of any data or information. Visual elements such as charts, graphs, and maps are someof the data visualization tools that provide the viewers with an easy and accessible way of understandingthe represented information. Time series is a sequence of data points in that follow the order in which they occurred. sequence, most often gathered in regular intervals. Time series analysis can be applied to any variable that changes over time and generally speaking, usually data points that are closer together are more similar than those further apart. In this paper the Covid-19 dataset is taken from the web scraping technique and then the process of EDA, Data Visualization and Time Series analysis is performed on it.And with the other dataset the text classification is performed using the natural language processing (NLP) techniques.

KEYWORDS: EDA (Exploratory Data Analysis), Data Visualization, Time Series Analysis, Web Scraping, Covid-19 Dataset, Text Classification, Natural language processing (NLP).

I. INTRODUCTION.

Data visualisation is the representation of the data in a graphical form. It converts the raw data present into a much more effective form, as a result this data can be communicated to the user more efficiently. Data Visualisation makes the data more understandable and user-friendly. In the age of big data and data analytics the huge amount of data present can be put to a great use with the help of data visualisation. Data visualisation opens up an ocean of opportunity in field of big data and data analytics. Data visualisation brings us a step closer to advance technologies such as Artificial Intelligence and Big Data. Due to the complexity in analysing rainfall and the ability to learn from the past datasets Data visualisation plays a vital role in the analysis of the large amount of data and it could prove a great aid to Artificial Intelligence. The data can be easily used with the help of tools and libraries such as seaborn, NumPy and pandas present in the Python language. These libraries consist of various tools which can be used convert the raw data into a more understandable form. In this paper bar and line graphs are used to represent the data. Visualising the data has great benefits. It gives a clearer view of thing by converting the raw data into a much more understandable and user-friendly one. By visualising the data can make sense of large data sets. Visualising also encourage drawing out patterns trends and comparing various data. Drawing out these patterns has great use in Search Engine Optimisation (SEO). Visualising the data shows all the intricacies of perplexing subjects which is of great use in the field of Artificial Intelligence. The datasets used in this paper is shared by Kaggle Repositories[1]. It contains the Covid-19 detail of all Continents and Countries.

Machine Learning algorithms learning is based on trial and error method quite opposite of conventional algorithms, which follows the programming instructions based on decision statements like if-else. One of the most significant areas of ML is forecasting, numerous standard ML algorithms have been used in this area to guide the future course of actions needed in many application areas including weather forecasting, disease forecasting, stock market forecasting as well as disease prognosis. Various regression and neural network models have wide applicability in predicting the conditions of patients in the future with a specific disease. There are lots of studies performed for the prediction of different diseases using machine learning techniques such as coronary artery disease, cardiovascular disease

prediction, and breast cancer prediction. In particular, the study is focused on live forecasting of COVID-19 confirmed cases and study is also focused on the forecast of COVID19 outbreak in India containing the details of different states and union territories. These prediction systems can be very helpful in decision making to handle the present scenario to guide early interventions to manage these diseases very effectively. This study aims to provide an early forecast model for the spread of novel coronavirus, also known as SARS-CoV-2, officially named as COVID-19 by the World Health Organization (WHO). COVID-19 is presently a very serious threat to human life all over the world. At the end of 2019, the virus was first identified in a city of China called Wuhan, when a large number of people developed symptoms like pneumonia. It has a diverse effect on the human body, including severe acute respiratory syndrome and multi-organ failure which can ultimately lead to death in a very short duration. Thousands of new people are reported to be positive every day from countries across the world. The virus spreads primarily through close physical contact from one person to another person, by respiratory droplets, or by touching the contaminated surfaces. The most challenging aspect of its spread is that a person can possess the virus for many days without showing symptoms. The causes of its spread and considering its danger, almost all the countries have declared either partial or strict lockdowns throughout the affected regions and cities. Medical researchers throughout the globe are currently involved to discover an appropriate vaccine and medications for the disease. Since there is no approved medication till now for killing the virus so the governments of all countries are focusing on the precautions which can stop the spread. Out of all precautions, "be informed" about all the aspects of COVID-19 is considered extremely important. To contribute to this aspect of information, numerous researchers are studying the different dimensions of the pandemic and produce the results to help humanity. To contribute to the current human crisis our attempt in this study is to develop a forecasting system for COVID-19. The forecasting is done for the three important variables of the disease for the coming 10 days: 1) the number 0f New confirmed cases. 2) the number of death cases 3) the number of recoveries.

Text classification (a.k.a. text categorization or text tagging) is the task of assigning a set of predefined categories to open-ended. Text classifiers can be used to organize, structure, and categorize pretty much any kind of text – from documents, medical studies and files, and all over the web. The text classification involves the coronavirus related tweet dataset.

II. RELATED WORK.

A. Exploratory Data Analysis on Covid-19 Dataset.

Exploratory Data Analysis (EDA) in Python is the first step in your data analysis process developed by "**John Tukey**" in the 1970s. In statistics, exploratory data analysis is an approach to analysing data sets to summarize their main characteristics, often with visual methods. By the name itself, we can get to know that it is a step in which we need to explore the data set.

By completing the **Exploratory Data Analysis,** you will have many plots, heat-maps, frequency distribution, graphs, correlation matrix along with the hypothesis by which any individual can understand what your data is all about and what insights you got from exploring your data set.Steps in Exploratory Data Analysis In Python, 1] Description of data,2] Handling missing data,3] Handling outliers,4] Understanding relationships and new insights through plots.

Description of data.

We need to know the different kinds of data and other statistics of our data before we can move on to the other steps. A good one is to start with the **describe ()** function in python. In Pandas, we can apply describe() on a Data Frame which helps in generating descriptive statistics that summarize the central tendency, dispersion, and shape of a dataset's distribution, excluding Nan values.

Handling missing data.

Data in the real-world are rarely clean and homogeneous. Data can either be missing during data extraction or collection due to several reasons. Missing values need to be handled carefully because they reduce the quality of any of our performance matrix. It can also lead to wrong prediction or classification and can also cause a high bias for any given model being used. There are several options for handling missing values. However, the choice of what should be done is largely dependent on the nature of our data and the missing values. Below are some of the techniques:

- Drop NULL or missing values: This is the fastest and easiest step to handle missing values. However, it is not generally advised. This method reduces the quality of our model as it reduces sample size because it works by

International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 9, pp: 694-701 www.ijaem.net ISSN: 2395-5252

deleting all other observations where any of the variables is missing.

- Fill Missing Values: This is the most common method of handling missing values. This is a process whereby missing values are replaced with a test statistic like mean, median or mode of the particular feature the missing value belongs to.
- Predict Missing values with an ML Algorithm: This is by far one of the best and most efficient methods for handling missing data. Depending on the class of data that is missing, one can either use a regression or classification model to predict missing data.

Handling outliers.

An outlier is something which is separate or different from the crowd. Outliers can be a result of a mistake during data collection or it can be just an indication of variance in your data. Some of the methods for detecting and handling outliers.

Boxplot: A box plot is a method for graphically depicting groups of numerical data through their quartiles. The box extends from the Q1 to Q3 quartile values of the data, with a line at the median (Q2). The whiskers extend from the edges of the box to show the range of the data. Outlier points are those past the end of the whiskers. Boxplots show robust measures of location and spread as well as providing information about symmetry and outliers.

Scatterplot: A scatter plot is a mathematical diagram using Cartesian coordinates to display values for two variables for a set of data. The data are displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the vertical axis.

Z-score: The Z-score is the signed number of standard deviations by which the value of an observation or data point is above the mean value of what is being observed or measured. While calculating the Z-score we re-scale and centre the data and look for data points that are too far from zero. These data points which are way too far from zero will be treated as the outliers. In most of the cases a threshold of 3 or -3 is used i.e. if the Z-score value is greater than or less than 3 or -3 respectively, that data point will be identified as outliers.

IQR: The interquartile range (IQR) is a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles.

IQR = Q3 − Q1.

Understanding relationships and new insights through plots.

We can get many relations in our data by visualizing our dataset. Let's go through some techniques in order to see the insights.

Histogram: A histogram is a great tool for quickly assessing a probability distribution that is easy for interpretation by almost any audience. Python offers a handful of different options for building and plotting histograms.

Heatmaps: The Heat Map procedure shows the distribution of a quantitative variable over all combinations of 2 categorical factors. If one of the 2 factors represents time, then the evolution of the variable can be easily viewed using the map. A gradient colour scale is used to represent the values of the quantitative variable. The correlation between two random variables is a number that runs from -1 through 0 to +1 and indicates a strong inverse relationship, no relationship, and a strong direct relationship, respectively.

The Tools Exploratory Data Analysis

In programming, we can accomplish EDA using Python. Some of the important packages in Python are:1.Pandas,2.Numpy,3.Matplotlib,4.Seabornetc

B. Time Series Analysis on Covid-19 Dataset.

Steps to perform the Time Series analysis on the dataset are:

Step 1 — Installing Packages: The pandas library, which offers a lot of flexibility when manipulating data, and the statsmodels library, which allows us to perform statistical computing in Python. Used together, these two libraries extend Python to offer

International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 9, pp: 694-701 www.ijaem.net ISSN: 2395-5252

greater functionality and significantly increase our analytical toolkit.

Step 2 — Loading Time-series Data

Step 3 — Indexing with Time-series Data: You may have noticed that the dates have been set as the index of our pandas Data Frame. When working with time-series data in Python we should ensure that dates are used as an index, so make sure to always check for that.

Step 4 — Handling Missing Values in Time-series Data: Real world data tends be messy. As we can see from the plot, it is not uncommon for time-series data to contain missing values.

Step 5 — Visualizing Time-series Data: When working with time-series data, a lot can be revealed through visualizing it. A few things to look out for are:1] **seasonality**: does the data display a clear periodic pattern? 2] **trend**: does the data follow a consistent upwards or downward slope? 3] **noise**: are there any outlier points or missing values that are not consistent with the rest of the data?

C. Coronavirus Tweets NLP – Text Classification.

Sentiment analysis aims to estimate the **sentiment polarity** of a body of text based solely on its content. The sentiment polarity of text can be defined as a value that says whether the expressed opinion is **positive** (polarity=**1**), **negative** (polarity=**0**), or neutral. In this tutorial, we will assume that texts are either positive or negative, but that they can't be neutral. Under this assumption, sentiment analysis can be expressed as the following classification problem: **Feature**: the string representing the input **text, Target**: the text's **polarity** (0 or 1).

We need to transform the main feature — i.e., a succession of words, spaces, punctuation and sometimes other things like emojis — into some numerical features that can be used in a learning algorithm. To achieve this, we will follow two basic steps:1] A **pre-processing** step to make the texts cleaner and easier to process, 2] And a **vectorization** step to transform these texts into numerical vectors.

Pre-Processing: A simple approach is to assume that the smallest unit of information in a text is the word (as opposed to the character). Therefore, we will be representing our texts as **word sequences**.

Vectorization: After getting the word sequences by performing pre-processingneed a way to transform these word sequences into numerical features: this is **vectorization**.The simplest text vectorization technique is Bag of Words (BOW). It starts with a list of words called the vocabulary (this is often all the words that occur in the training data). Then, given an input text, it outputs a numerical vector which is simply the vector of word counts for each word of the vocabulary. Using BOW is making the assumption that the more a word appears in a text, the more it is representative of its meaning. Therefore, we assume that given a set of positive and negative text, a good classifier will be able to detect patterns in word distributions and learn to predict the sentiment of a text based on which words occur and how many times they do.

To use BOW vectorization in Python, we can rely on Count Vectorizer from the scikit-learn library. In addition to performing vectorization, it will also allow us to remove stop words (i.e., very common words that don't have a lot of meaning, like this, that, or the). scikit-learn has a built-in list of stop words that can be ignored by passing stop_words="english" to the vectorizer.

III. PROBLEM STATEMENT AND DATA.

Problem Statement 01: The Exploratory data analysis on COVID-19 19 dataset.

The dataset large number of rows, each containing Sno, Date, Time, State or Union Territories, Confirmed Indian National Cases, Confirmed Foreign National Cases, cured cases, death cases and Total confirmed cases related to Coronavirus or Covid19. Our objective is to use this data, explore it, and generate insights from it, and also visualizing the data using Seaborn and Matplot libraries.

The dataset is taken from the website named Kaggle, it has about 7085 rows and 9 columns as named above.

Solution for problem statement 01: Performing the data analysis, data pre-processing, data visualization and fit the various machine learning algorithms on the training dataset to predict their scores.

Results of Problem Statement 01:

1] Describe the dataset using the Python command.

```
datasetname. describe ()
```

and the complete information of the dataset is obtained from the command

```
datasetname.info ()
```

International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 9, pp: 694-701 www.ijaem.net ISSN: 2395-5252

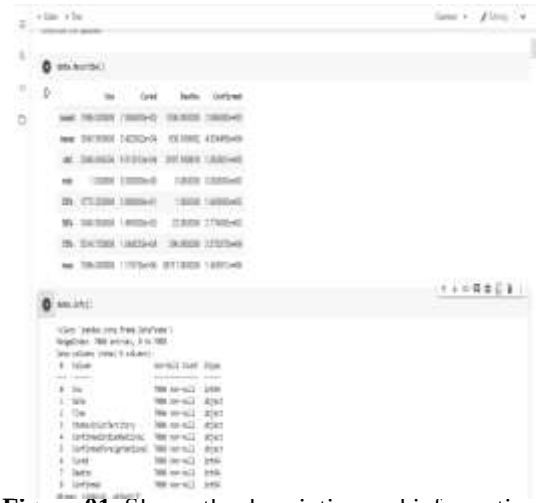

Figure 01: Shows the description and information on the dataset.

2] Data Visualization.

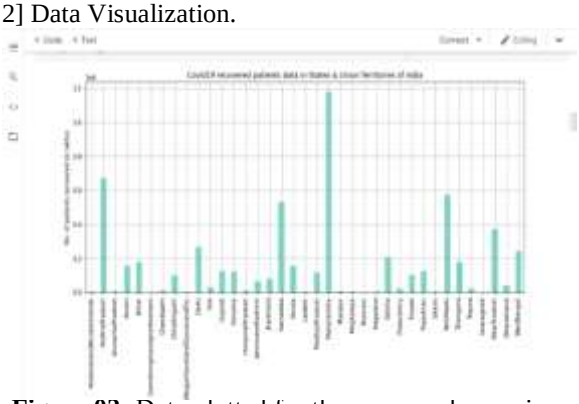

Figure 02: Data plotted for the recovered cases in states and union territories.

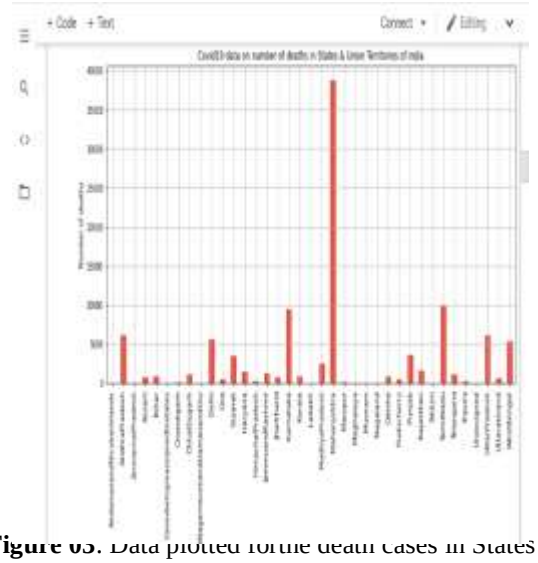

Figure 03: Data plotted for the death cases in States and Union territories.

3] Handling the missing values.

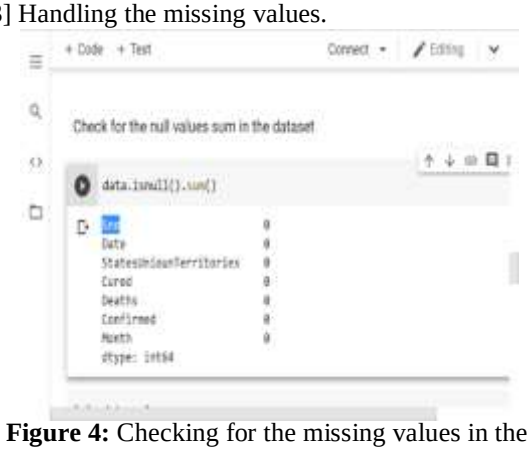

Figure 4: Checking for the missing values in the dataset.

Firstly the sum of the null values is checked as shown in the figure 4 , as sum of missing values is zero for all the columns the handling of missing values is not performed.

4] Split the dataset into Features and Labels and then train the data for different machine learning algorithms.

Figure 05: The Machine learning regression models that are fit on the training data and score of machine learning data is checked on the test data.

Problem Statement 02: Performing Time Series Analysis for the Covid19 dataset.

Examine how the changes associated with chooses data point compares it to shifts in other variables over the same time period.

Solution of Problem Statement 2: Perform the time series analysis by performing Dickey-Fuller test to check stationarity of data and to reject null hypothesis, ARIMA model, Trend analysis using the log transformation

Results of Problem Statement 2:

International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 9, pp: 694-701 www.ijaem.net ISSN: 2395-5252

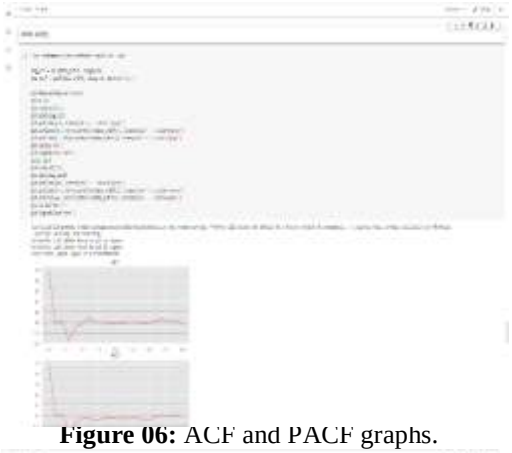

Figure 06: ACF and PACF graphs.

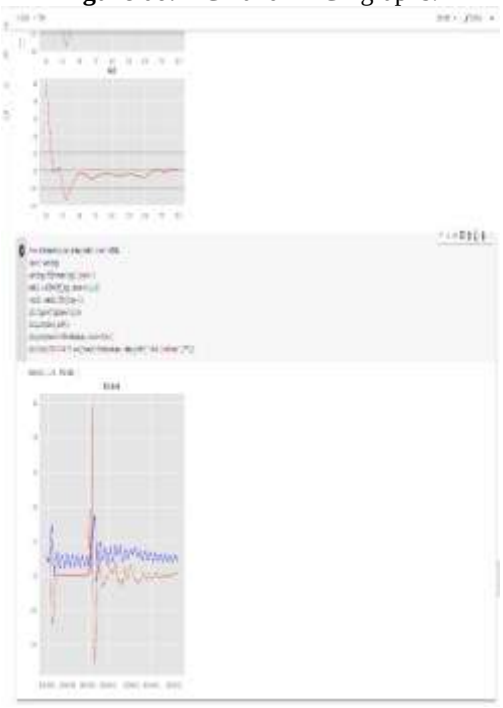

Figure 07: ARIMA model.

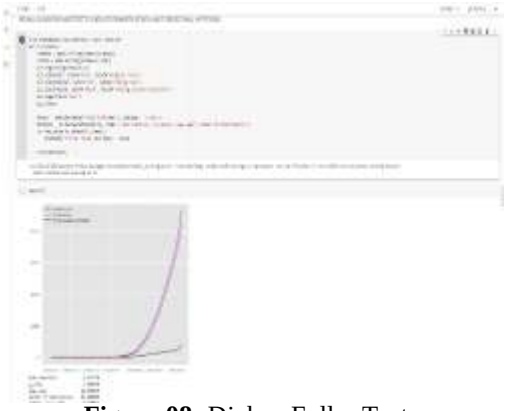

Figure 08: Dickey Fuller Test.

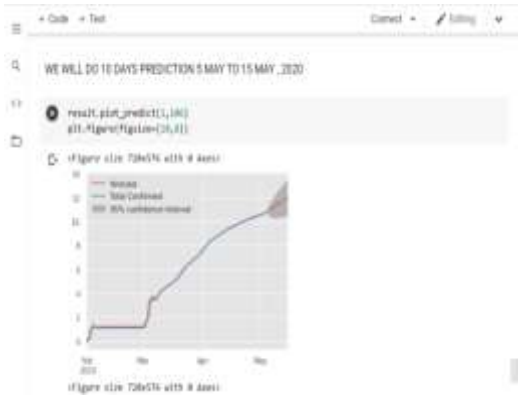

Figure 09: 10 days prediction of the dataset.

Problem Statement 03: Performing the Text Classification using the NLP techniques.

Examine the Sentiments in the coronavirus related tweets using the Natural language processing techniques.

Solution of problem statement 03: Here the dataset containing the tweets related to coronavirus is taken and then is performed with the text classification using the nlp techniques that classify the sentiments of the text , the word cloud is build to for positive and the negative words of the tweets, and also the sequential neural network is build ,and the classification report is obtained containing the precision, recall, f1-score and support results.

Results of Problem statement 03:

Figure 10: Word cloud that mentions the negative sentiments in training and testing data.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 9, pp: 694-701 www.ijaem.net ISSN: 2395-5252

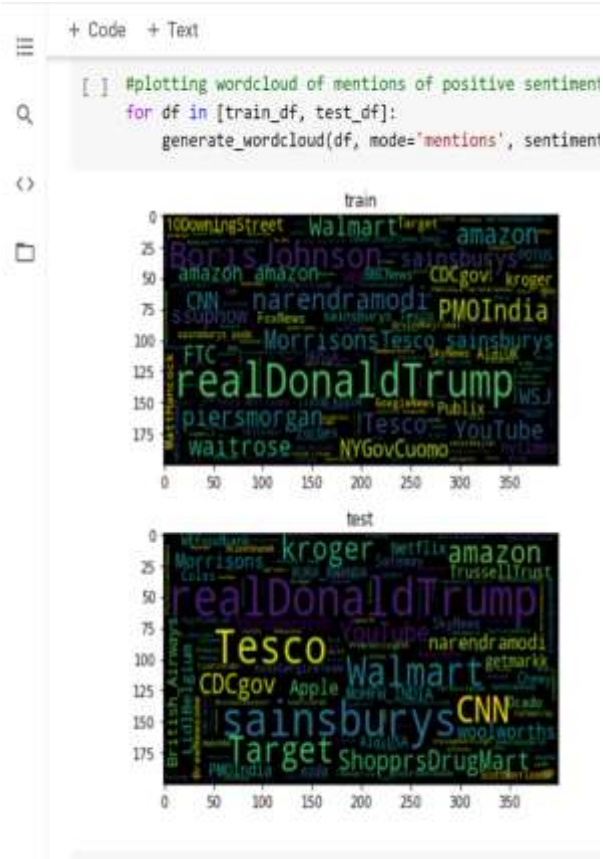

Figure 11: Word cloud that mentions the positive sentiments in training and testing data.

Figure 12: Distribution of sentiments in data.

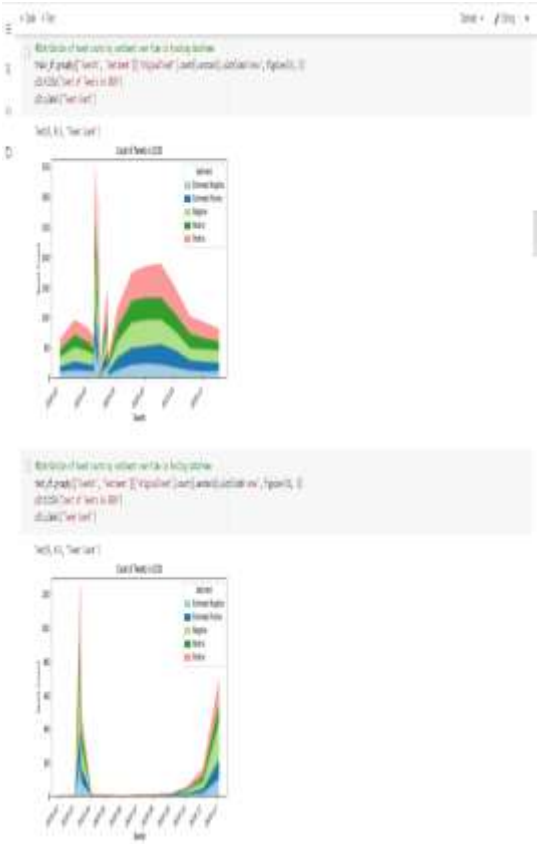

Figure 13: Distribution of tweet counts by sentiments over time in training and testing data.

Figure 14: Building Sequential neural network using TensorFlow.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 9, pp: 694-701 www.ijaem.net ISSN: 2395-5252

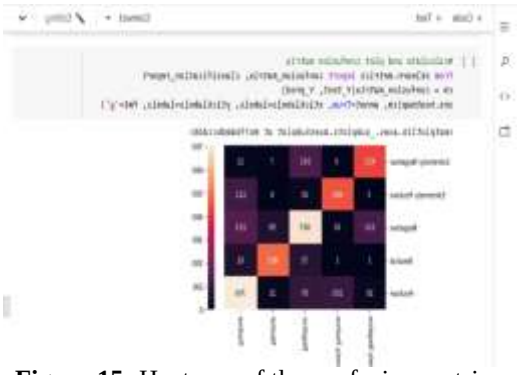

Figure 15: Heatmap of the confusion matrix.

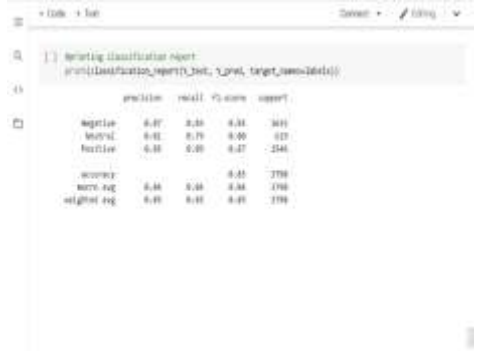

Figure 16: Classification report

IV. CONCLUSION.

The results obtained in this paper can be used for the prediction of Covid19 in India with the help of regression which can be of great benefit , and also the sentiments of the tweets helps in capturing the sentiments of the people regarding coronavirus, Time series analysis of the dataset would be a great benefit to analyse the when and where the Covid19 outburst is more and helps to prevent it in the future.

V. ACKNOWLEDGMENT

This research contribution is a part of Postgraduate research and content development at institute of Computer Science Engineering.

REFERENCES.

[1]. Datasets for Covid19 or Coronavirus available in https://www.kaggle.com/datasets