**GROWTH AND SENTIMNT ANALYSIS ON CORONAVIRUS**

Project Report

Submitted by

**Anmol Kumar Tanmay Ambegaokar**

170906482 170906292

Under the guidance of

**SATYAKAM**

Assistant Professor – Senior Scale

Department of Electrical &Electronics,

MIT MAHE, Manipal

in partial fulfilment of the requirements for the award of the degree of

# **BACHELOR OF TECHNOLOGY**

# **IN**

# **ELECTRICAL AND ELECTRONICS ENGINEERING**

# 

# **DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING**

**MANIPAL INSTUTUTE OF TECHNOLOGY**

(A Constituent Unit of Manipal Academy of Higher Education)

MANIPAL-576104, KARNATAKA, INDIA

June 2021

A screenshot of a cell phone

Description generated with very high confidence**DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING**

Manipal

15/06/2021

**CERTIFICATE**

This is to certify that the project titled **GROWTH AND SENTIMENT ANALYSIS ON CORONAVIRUS** is a record of the bonafide work done by Anmol Kumar (*Reg. No. 170906482*) and Tanmay Ambegaokar (*Reg. No. 170906292*) submitted in partial fulfilment of the requirements for the award of the Degree of Bachelor of Technology (B.Tech.) in **ELCTRICAL AND ELECTRONICS ENGINEERING** of Manipal Institute of Technology, Manipal, Karnataka, (A Constituent Institute of Manipal Academy of Higher Education), during the academic year 2020-2021.

|  |  |
| --- | --- |
| **SATYAKAM**  *Assistant Professor – Senior Scale**Department of Electrical & Electronics,*  *M.I.T., MANIPAL* | **Prof. Dr. Chandrashekara S Adiga** *HOD, Dept of E&E,*  *M.I.T., MANIPAL* |

**AKNOWLEDGMENTS**

We would like to take this opportunity to thank the Director of our college, Dr. D Srikanth Rao and the HOD of our department, Dr Chandrashekara S Adiga, for giving us this wonderful opportunity to work on this project and express our knowledge.

We would also like to extend our gratitude towards our Project coordinator, Dr Vijayalaxmi, for her continuous guidance and support of our efforts while we were working on our project and for being very approachable whenever we needed her counsel.

We would further like to acknowledge our project guide, Mr. Satyakam, for his continued support, guidance and help at each step of the project from start to end. He was regularly present in our discussions and help us work in a systematic manner and motivated us to produce the best work we can during the difficult times of Covid-19 pandemic.

**ABSTRACT**

There is no shortage of data in today’s world. If processed, analysed and modelled properly through right methodology, a variety of results, predictions and inferences can be generated using these data sets. Such a study could be very beneficial in understanding the growth and behaviour of the Covid-19 pandemic. Coronavirus was first detected in February 2003 but was contained by July 2003. The present version of Coronavirus outbreak took place in late 2019 and soon spread to many countries across the globe. These countries tried various methods to combat the pandemic. Predicting the trend of spread and growth of this virus could be helpful to tackle it. Further, a comprehensive data analysis is also done on the sentiments and general reaction of people in India with respect to the in -house vaccine prepared in early 2021. Based on data extracted from tweets by filtering on the basis hashtag ‘#covaxin’ and ‘#covishield’, sentiments of people on vaccine is analysed.

The datasets of Covid-19 cases from India is used for the analysis of the behaviour of their growth. The data set of SARS outbreak of 2003 is also taken with similar fields to understand the relation between the two outbreaks occurring at two different times. To perform growth analysis, we have chosen three models- linear regression, logistic regression, Gompertz curve model. After training and testing the models, the accuracy and the error rate have been evaluated for all the three models to find out which model is best suited to predict the growth of coronavirus cases for both the datasets. For Sentiment Analysis, raw data has been collected from Twitter API and then pre-processed[1]. This processed data has been labelled into positive, negative, and neutral sentiments using VADER tool. The data is plotted for both the vaccines and for January, February and March separately.

Once the SARS and Covid-19 datasets have been modelled and evaluated, it was found that Logistic Regression gave the best prediction with maximum accuracy and least error rate. For the sentiment analysis, there was a higher percentage of positive views for Covaxin vaccine than Covishield Vaccine. There was an increase in percentage of positive views for vaccines from January to February but, it decreased from February to March.

The possible factors for the results in sentiment analysis are the different efficacy results of the two vaccines and start of the second phase of vaccination in March. We have used the platform Google Collab and Jupyter Notebook to execute our codes. The codes have been written in Python language. Some of the packages and libraries imported for writing the codes for growth analysis were Pandas, Numpy, and sklearn. For sentiment analysis, we have used nltk, VADER, and tweepy. For visualization and plotting of graphs, we have used matplotlib and seaborn packages.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **TABLE OF CONTENTS** | | | | | | | |
|  | | | | | | | Page No |
| Acknowledgement | | | | | |  | i |
| Abstract | | | | | |  | ii |
| List of Tables | | | | | |  | iv |
| List Of Figures | | | | | |  | v |
| **Chapter 1** | | | | | **INTRODUCTION** | | **1** |
|  | | **1.1** | | Introduction | | | 1 |
|  | | **1.2** | | Motivation | | | 1 |
|  | | **1.3** | | Relevance and Impact of the work | | | 2 |
|  | | | | | | | |
| **Chapter 2** | | | | | LITERATURE REVIEW & OBJECTIVES | | **3** |
|  | | **2.1** | | Literature Review | | | 3 |
|  | | **2.2** | | Objectives | | | 5 |
|  | | | | | | | |
| **Chapter 3** | | | | | METHODOLOGY | | **7** |
|  | 3.1 | | Growth analysis | | | | 7 |
|  | 3.2 | | Sentiment analysis | | | | 13 |
|  | | | | | | |  |
| **Chapter 4** | | | | | **CONTRIBUTION OF EACH STUDENT** | | **18** |
|  | | **4.1** | | Contribution of Anmol Kumar | | | 18 |
|  | | **4.2** | | Contribution of Tanmay Ambegaokar | | | 20 |
|  | | | | | | | |
| **Chapter 5** | | | | | **RESULT ANALYSIS** | | **22** |
|  | | **5.1** | | Growth Analysis | | | 22 |
|  | | **5.2** | | Sentiment Analysis | | | 27 |
|  | | | | | | | |
| **Chapter 6** | | | | | **CONCLUSION AND FUTURE SCOPE** | | **30** |
|  | | **6.1** | | Growth Analysis | | | 30 |
|  | | **6.2** | | Sentiment Analysis | | | 31 |
|  | | | | | | | |
| **REFERENCES** | | | | | | | **32** |
| **PROJECT DETAILS** | | | | | | | **33** |

**LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
| **Table No** | **Table Title** | **Page No** |
| 5.1 | Co-efficient of regression values for all models | 30 |
| 5.2 | Mean absolute Error for all models | 30 |
| 5.3 | Sentiment analysis results for Vaccines | 31 |
| 5.4 | Month wise sentiment analysis results for both Vaccines | 31 |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Figure No** | **Figure Title** | **Page No** |
| 3.1 | Unique countries affected by SARS 2003 | 8 |
| 3.2 | Plot of countries affected by SARS 2003 | 8 |
| 3.3 | Confirmed cases in India- state wise | 9 |
| 3.4 | Steps followed in Sentiment Analysis of tweets related to public reactions on vaccination drive | 12 |
| 3.5 | Stages of Twitter Data Preprocessing | 14 |
| 5.1 | Regression Line Distribution of Confirmed cases vs Deaths for SARS 2003 | 21 |
| 5.2 | Error Rate and R2 score | 22 |
| 5.3 | Regression Line Distribution of Confirmed cases vs Deaths for Covid-19 | 22 |
| 5.4 | Gompertz curve for SARS 2003 in Singapore | 23 |
| 5.5 | Curve fitting for SARS 2003 | 23 |
| 5.6 | Gompertz curve for Covid-19 in Kerala | 24 |
| 5.7 | Curve fitting for Covid-19 | 24 |
| 5.8 | Logistic Regression Curve on SARS 2003 | 25 |
| 5.9 | Logistic Regression Curve on Covid-19 | 25 |
| 5.10 | January Sentiments % pie chart | 26 |
| 5.11 | January Sentiments Frequency Bar plot | 26 |
| 5.12 | February Sentiments % pie chart | 26 |
| 5.13 | February Sentiments Frequency Bar plot | 26 |
| 5.14 | March Sentiments % pie chart | 27 |
| 5.15 | March Sentiments Frequency Bar plot | 27 |
| 5.16 | Covaxin Sentiments % pie chart | 27 |
| 5.17 | Covaxin Sentiments Frequency Bar plot | 27 |
| 5.18 | Covishield Sentiments % pie chart | 28 |
| 5.19 | Covishield Sentiments Frequency Bar plot | 28 |

**CHAPTER 1**

**INTRODUCTION**

* 1. *INTRODUCTION*

Data analysis is done on vast amount of raw data. The recent Covid-19 pandemic has provided with datasets in abundance. These datasets if used and processed properly could provide valuable insights to tackle the pandemic and help us be more equipped for any future scenarios. Several studies have been carried out using these datasets for prediction of the behaviour of the pandemic and behaviour of people in response to the pandemic.

Growth analysis help us understand the peak of the curve of the cases and gives us an approximate idea of how the ongoing rise in cases will result to in future. These results are obtained using the different models by training the datasets on them and then predicting. The second part of the project deals with the sentiment analysis of tweets related to opinions of people on the vaccination drive. Sentiment analysis is usually carried out for predicting the rise or fall of stocks using the opinions of people on twitter. Every individual has their own opinions and twitter is one of the leading websites where people come and express their opinion on any issue. Such was the case when vaccines were rolled out earlier this year. The severity of the situation led to fast tracking the clinical trials of vaccines and making them available consumption for public. This had led to scepticism and paranoia among the masses with regards to the legitimacy and effectiveness of the vaccine. Thus, this section of the project aims to get an idea of the general mood of the public for the two vaccines – Covishield and Covaxin separately and then during each month of January, February and March.

* 1. *MOTIVATION*

The motivation for taking this project up was to understand the relation between the spread of the pandemic between different regions (India and different countries) and between different time (2003 SARS outbreak and 2019 SARS outbreak). The datasets of Covid-19 basically comprise of date, number of cases, number of deaths and number of recovered cases along with the place. The datasets were taken from Kaggle. The data was large and had detailed information about the time of recording the Covid-19 case and the date. The SARS 2003 dataset had around 2000 rows. The Covid-19 in India dataset had around 9000 rows. These fields have led to studies being conducted on growth and spread in various countries. The motivation is also to test different types of models on these two datasets and see the accuracy on different parameters of linear and non-linear models and also within different non-linear models (Logistic regression and Gompertz Model). A good model will help determine when has the peak occurred and how long will the peak continue. This kind of analysis helps in predicting the growth and decline of a particular function with respect to another continuous function which in this case is date and time.

The motivation for taking up the sentiment analysis section of the project was to try and understand the general mood of the public when something so necessary like vaccines are fast tracked and the factors which may have affected the positive or negative sentiments. Previous works have been done majorly on initial few days of vaccine roll out. We have tried to look into a broader time frame keeping in mind both the vaccines.

* 1. *RELEVANCE AND IMPACT OF THE WORK*

Our project aims to analyse the data regarding the SARS 2003 outbreak and Covid-19 in India. We want to do a detailed analysis of both the datasets and conduct growth analysis. The growth analysis will give predictions with respect to how the number of have increased over time. It will also give us a fair idea on when the number of cases and deaths are expected to flatten out. Since, we have conducted our analysis using three different models, our work will tell us which model is best suited to analyse exponential data and which model gives the most accurate result.

We have also extracted data from twitter to conduct sentiment analysis on the vaccines namely, Covaxin and Covishield., being distributed in India. Getting an idea of the reactions should act as a good feedback to the efforts being carried out in bringing confidence in masses about the legitimacy of the vaccine. Thus, this analysis gives a good reality check of the efforts being put forth to eradicating the Covid-19 virus.

**CHAPTER 2**

**LITERATURE REVIEW & OBJECTIVES**

In this chapter, we discuss the different types of analysis done in this project namely – Growth Analysis and Sentiment Analysis. We also discuss the types of models implemented. Finally, we also mention the objectives of the project and the individual objectives of the student.

The project is divided into two parts – Growth Analysis and Sentiment Analysis. For the growth analysis, we have taken two datasets of Covid-19 in India and SARS 2003 outbreak in the world. We have performed linear and logistic regression on both these datasets. In addition to this, we have also implemented Gompertz growth model to predict how the cases will increase with time. For the Sentiment Analysis, we have extracted data from Tweets related to the vaccines distributed in India and the general reaction of the public towards it.

*2.1 Literature Review*

Regression is a technique which establishes a relationship with a dependent variable and an independent variable. It can be linear or non-linear. As Vikas Kumar Sharma, in his article on “Modelling and Forecasting Covid-19 Growth Curve in India” [2], rightly says that originally, when there was no lockdown, cases were rising exponentially with respect to time. Once, the government implemented the lockdown, the cases will at some point of time stop rising exponentially. Or in other terms, at some point, the curve will flatten. This kind of regression is called polynomial regression. It will help understand how long the cases will grow exponentially, and at what point will it start to decline. Thus, it is useful in predicting the growth of covid-19 in India using regression.

Kathleen M,.C Tjorve and Evan Tjorve, in their article on “ The use of Gompertz Model on Growth Analysis and new Gompertz Model Approach”[3] state that the Gompertz model is one of the most frequently used sigmoid models fitted to growth data and other data. They mention that researchers have fitted the Gompertz model on bird growth, animal growth and even tumor growth. Because of the extensive usage of the Gompertz model along with logistic regression, we have decided to use these models to model the coronavirus cases.

Social media sites have become a breeding ground for disinformation. Twitter is one such platform where people share their views. Some of these views are often in the form of information and due to lack of fact checking, it is very easy to spread disinformation among the users. Sometime, this often leads to paranoia and has a negative effect. Before the 2016 US elections, more than 6.6 million tweets were linked to fake and conspiracy news publishers[4]. This has been a recurring theme for some time now. Twitter recently had to intervene during the 2020 US elections when its President, Donald trump, was spreading fake news.

The Covid-19 pandemic in the early 2020 provided another such instance in disinformation was spread on platforms like Twitter, Instagram, Facebook and WhatsApp. "We’re not just fighting an epidemic; we’re fighting an infodemic”, were the words of WHO Director-General Tedros Adhanom Ghebreyesus at the Munich Security Conference on 15 February,2020[5]. Disinformation during that time created fear, drove societal disaccord and also proved to be harmful as seen in the case in Anand Vihar terminal in New Delhi where many migrant workers accumulated in the train station when a fake news regarding trains being functional went viral during lockdown. The study titled “An Exploratory Study on Covid-19misinformation on twitter” defines such misinformation, takes tweets and performs exploratory data analysis on the language distribution of tweets with disinformation, bot detection and hashtags used for the same. There is also a plot of sentiments like positive emotions, anxiety, anger, sadness and negative emotions created by the use of LICW method [6].

During the pandemic people have had different sentiments and the paper “Topic based Sentiment Analysis for COVID-19 Tweets “specifically focuses on twitter and extracts the most discussed topics during and after the first wave of the Coronavirus pandemic [7]. The eight sentiments were selected from each of the categories: neutral, positive, and negative. This helped in understanding the general mood of the people during the pandemic.

Similarly, when the vaccines were rolled out in the early 2020, there were a lot of questions marks on the effectiveness of the vaccine. People had both positive and negative sentiments on it and they took it to twitter to express them [8]. All the papers discussed earlier have not discussed on this aspect of the pandemic and there is no paper currently present on it. India had rolled out two vaccines – Covishield and Covaxin. The largest vaccination drive in the world started in India on 16th January, 2021. During the first phase, the healthcare workers and frontline workers were given the doses but quite a lot of people decided not take it. People took to twitter to express their distrust in the vaccine. An analysis of these tweets can help us understand the factors and the sentiments of people and role of right information dissemination.

*2.2 Objectives*

The main objective of our project is to find relevant data from Kaggle, clean the data and perform exploratory data analysis. The data is then presented in a visual form so that it can be better understood. Lastly, we aim to train the dataset and test it using logistic regression to find the best model which will give the best prediction regarding the growth of the novel coronavirus.

In January 2021, the Indian government started the largest vaccination drive in the world by starting vaccination for the healthcare workers. This was met by a lot of anxiety and nervousness among the general masses about the efficacy of the vaccine given that it had been approved in a short period of time. A lot of opinions were made on social media regarding the same. We will be performing sentimental analysis by using tweets from twitter as our data set to find out the general sentiments and opinion of the public

Objectives of Anmol Kumar

* To find the relevant datasets on SARS 2003 which has detailed information regarding the number of cases, number of deaths, number of recovered cases etc.
* To Apply for a twitter developer account and then using python libraries and packages to extract tweets with #covaxin from the twitter API and then converting them into datasets in the form of csv files.
* To pre-process the data collected from twitter using the nltk python package to make it ready for analysis.
* To perform Linear Regression on SARS 2003 and Covid-19 dataset.
* To perform Sentiment Analysis using the VADER tool
* To analyse the results of sentiment analysis and growth analysis.

Objectives of Tanmay Ambegaokar:

* To find relevant datasets on Covid-19 which has detailed information regarding the number of cases, number of deaths, number of recovered cases etc.
* To clean the data which involves dropping unnecessary columns and replacing null values
* To do the visualization of the data which involves plotting graphs.
* To perform Logistic Regression and Gompertz Curve Model and evaluate the model to find accuracy and error rate of the designed model.
* To extract tweets with #covaxin and #covishield from the twitter API and then converting them into datasets in the form of csv files.
* To analyse the results of sentiment analysis and growth analysis.

**CHAPTER 3**

**METHODOLOGY**

The project is divided into two parts namely:

* 1. Growth analysis using three machine learning algorithms on
* COVID-19 pandemic in India and
* SARS 2003 disease in different regions of the world
  1. Sentiment analysis on the vaccines of COVID-19 virus
  2. *Growth analysis*

For any data science project, there are mainly 5 steps which one needs to follow to achieve a good result. These steps are data collection, data pre-processing, exploring the data, model the data through machine learning algorithms and finally evaluating the model to get its accuracy and error rate.

1. Data Collection

For the growth analysis, the first step is to search and identify the datasets with relevant data fields necessary for analysis. The data fields which we identified for our analysis were date, time, confirmed number of cases, confirmed number of deaths, confirmed number of recovered cases. The idea behind identifying these data fields was to make a model which will predict how the cases will grow over time. The data was collected after looking extensively into the government websites like data.gov.in, John Hopkins University database and privately maintained reliable sources such as Kaggle. Eventually the following datasets were selected from Kaggle:

1. SARS 2003 OUTBREAK DATASET [1]
2. COVID 19 INDIA DATASET [2]
3. Data Pre-Processing

The next step is to clean the data for use in the analysis. For each dataset:

1. SARS 2003 OUTBREAK DATASET – The dataset had columns of Date, Country, Cumulative number, Number of deaths and Numbers recovered. This dataset has 5 columns and 2538 rows. To check if there are no null values after cleaning the data, we use .info() function. We found out that there are 0 null values.
2. Covid-19 India Dataset- The dataset had columns of S.no, Date, Time, State/Union Territory, Confirmed Indian Nationals, Confirmed Foreign Nationals, Cured, Deaths and Confirmed. This dataset has 9 columns and 9291 rows. For our analysis, we did not need the Confirmed Indian Nationals, Confirmed Foreign Nationals and time column and thus, were dropped from the dataset using .drop() function. After using the .info() function, we found out that there are 0 null values.

Once the cleaning step is implemented, only then can we move on to any kind of analysis.

1. Exploring the data

The next step is called exploratory data analysis. In this step, we explore the dataset and then plot it in a visual form so that it is better to understand the data and looks neat.

Before we start our analysis, it is important to import some basic libraries in python for better computation. Hence, we have imported pandas, numpy, matplotlib.plotly and seaborn libraries in our code.

1. Pandas - It is library which can be imported in python for data manipulation and analysis. It gives data structures and operations for analysis on numeric tables and time series.
2. Numpy – It provides a n-dimensional array. This makes it easy for data analysis on large number of rows and columns.
3. Matplotlib – It is used to create plots and graphs for better visualization
4. Seaborn – This library is used for making statistical graphs and used as a data visualization library. Scatterplots are made through this package.
5. Sklearn – This library is used to implement all the machine learning algorithms like regression, clustering. We can import any type of algorithm through this package

First, we look at the SARS 2003 dataset –

* In our dataset, there are many repeating countries which reported new cases on a different date. To help group all the data together, it was necessary to find out unique values in the country column. For this, we used unique() function on the column named “Countries”. This function returned an array with all the unique names of the country which had reported SARS 2003 Virus. We found out that there were 37 unique countries affected by SARS.

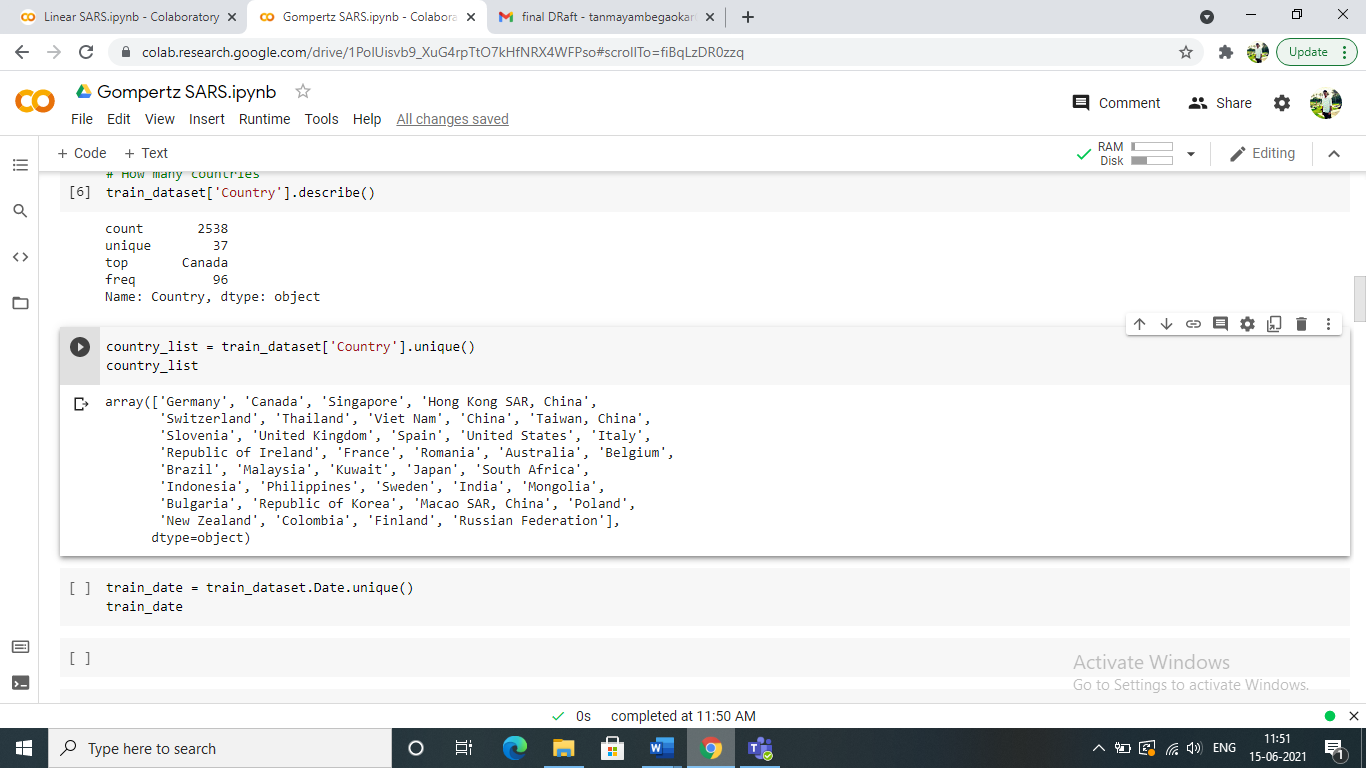


Fig 3.1. Unique countries affected by SARS 2003

* To plot all the countries and their respective total number of cases, we are going to make use of matplotlib library to plot the graph. matplotlib.pyplot provides a MATLAB-like way of plotting.

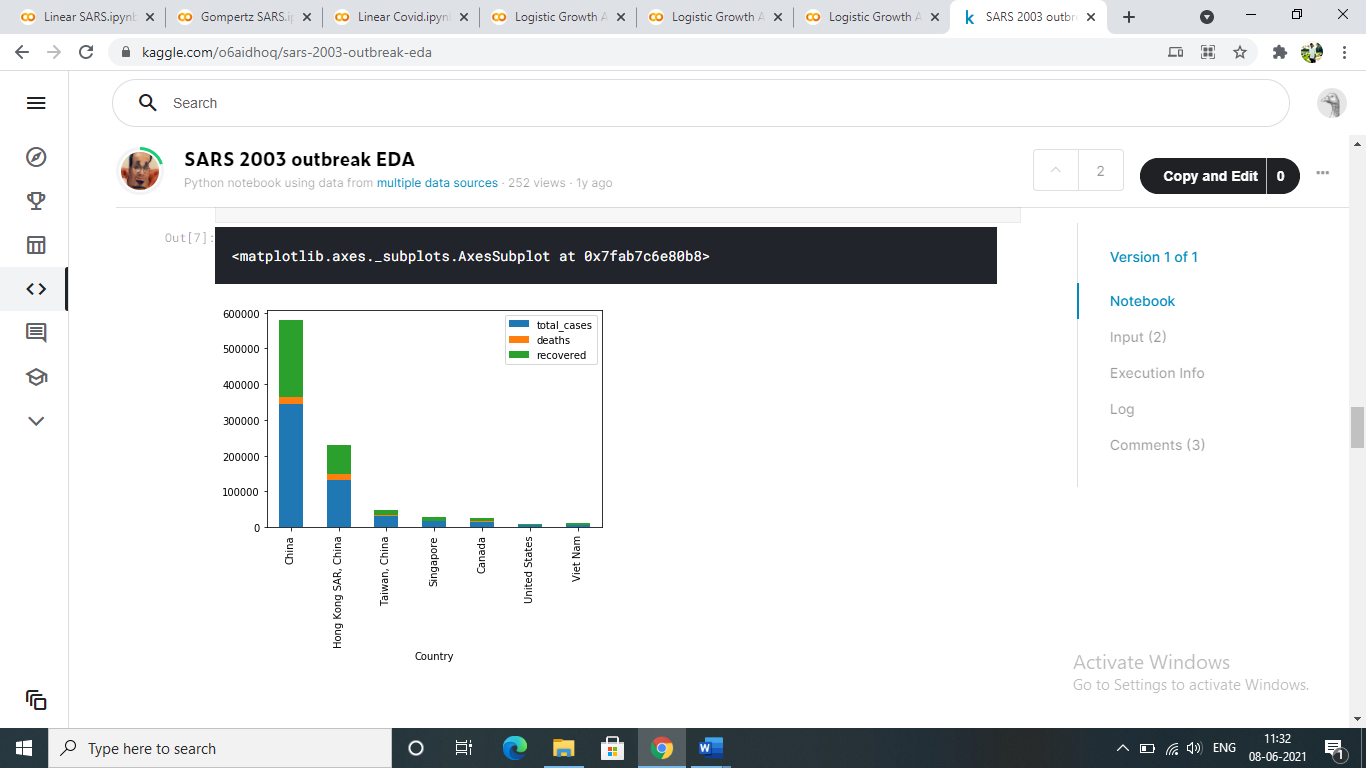


Fig 3.2. Plot of countries affected by SARS 2003

A similar analysis was also done on the Covid-19India dataset

* Like we found out the unique values of “country” column in SARS dataset, we did a similar .unique() function to find out unique values of “state” column. This column also involves cases from the Union Territories.
* Lastly, the total cases were plotted on a graph using plt.show() function. The cases for each day for each state were grouped together. Next, we did a summation to get the total confirmed cases for each state.

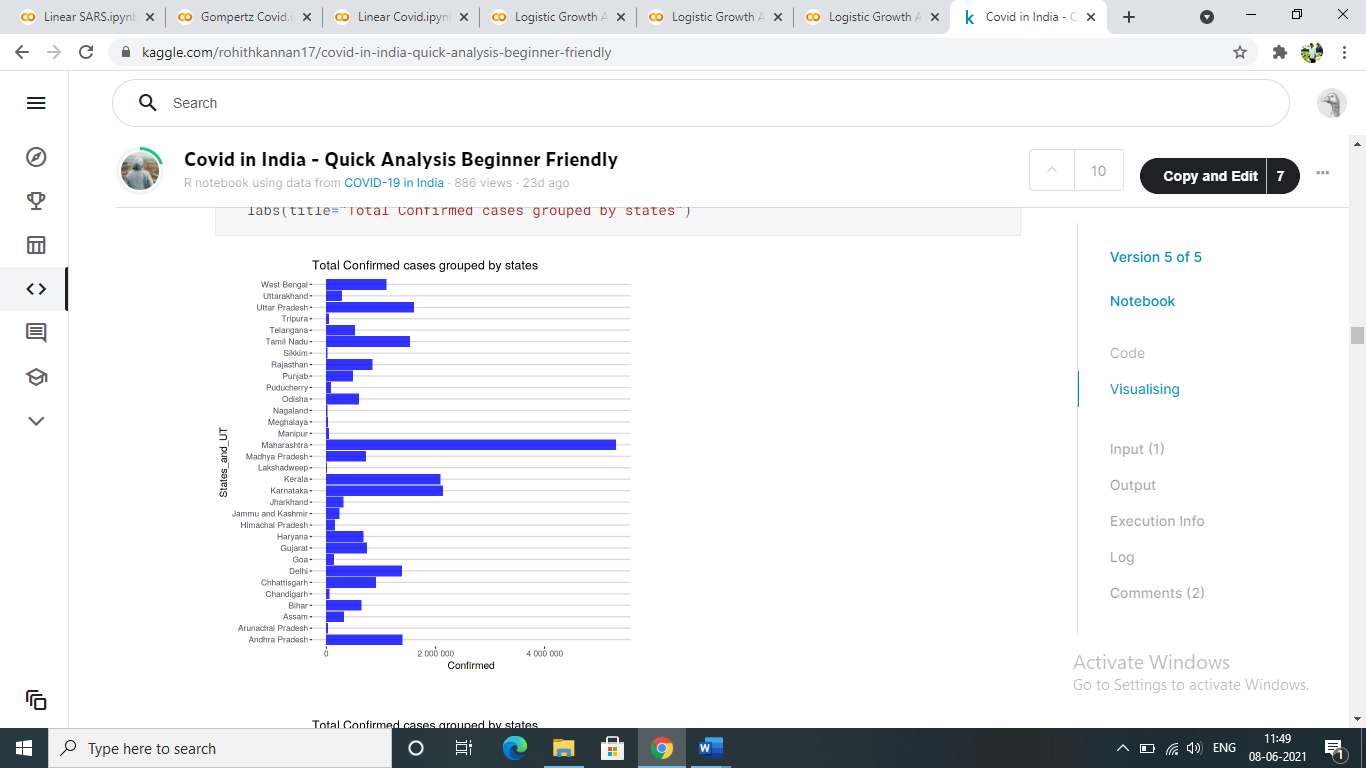


Fig 3.3. Confirmed cases in India- state wise

1. Modelling the Data

Finally, after all the data pre-processing and initial data exploration, the dataset is ready to be modelled using machine learning algorithms. We have chosen three machine learning algorithms for the prediction. We will be doing Linear regression, Logistic Regression and Gompertz Growth Model on both our datasets to get the best possible prediction.

i. Linear Regression

It is the relationship between 2 variables x and y. Usually, the x variable is the input or the independent variable whereas the other variable is the dependent one. The relationship is shown in the form of a straight line which tries to fit each variable. The mathematical expression for the regression line is y = mx + c, where c is the constant of regression.

In our project, the data is plotted using a scatter plot and a line is drawn which passes the scatter plot point. It tries to fit maximum data points and thus this line is also called best fit line.

We used the scatter plot, imported from the seaborn package to plot the graph. On the x-axis, we have the confirmed cases. And on the y-axis, we have the deaths recorded. A regression line is plotted through these data points using regplot() function. Once, the regression line is drawn, the dataset is split into testing and training split for model evaluation. The testing and testing are split randomly in a 30-70 split using the sklearn library to recognize trends and assess the model accuracy.

ii. Gompertz Growth Model

Gompertz Curve is a mathematical model used to plot a time-series model. It is a sigmoid function which describes growth as slowest during the start and end of a given time period. Thus, we chose Gompertz curve as one of the model for our prediction. The Covid-19 cases and the SARS disease grows slowly at the beginning before flattening out in the end. In the middle, it follows a non-linear path. The mathematical expression for the Gompertz curve is

Where, is an asymptote.

is the displacement along the x-axis.

is the growth rate.

Unlike the Linear Regression where we, split the dataset randomly in a 30-70 format, in this case, we have trained the dates of the start of the time period and the testing has been done on the latter dates. The training dataset recognizes patterns and the testing dataset is used to accurately predict the outcome. Hence, we use initial trends of the Covid-19 and SARS 2003 disease to predict how it will grow in the future.

To plot the gompertz curve, we have chosen a random state of India in the case of Covid-19 and a random country from the SARS dataset to showcase the growth curve. Lastly, curve fitting is done on the curve to provide a best fit to the specific dataset.

iii. Logistic Regression

Logistic Regression is the relationship between categorical data and dependent variables. Like the Gompertz curve, logistic curve is also a non-linear regression curve with a S-shaped curve. The cases will grow slowly at the start and then grow exponentially till flattening out in the end. The mathematical expression for the logistic curve is

Where, is the maximum value

is the logistic growth rate

is the number

is the value of the sigmoid

With this formula, we have predicted how the confirmed cases, number of deaths and number of recovered patients will grow over time.

1. Evaluating the model

To evaluate our model, we will be using three metrics to check how accurate our model is.

* Mean Absolute Error- The mean of the distance of all the data points from the best fit line is called mean absolute error. If it is closer to 0, it means all the points fit on the line perfectly.
* R2 score- This is the regression sum of square divided by the total sum of square. If the value is closer to 1, it means the model is more accurate.
* Mean Squared Error – The square of MAE. If it is closer to 0, it means the model is more accurate.
  1. *Sentiment analysis on tweets relating to the vaccine for Covid-19 virus*

The sentiment analysis was carried out on tweets collected in csv form from the twitter to understand the general mood of the masses in India when the two vaccines – Covishield and Covaxin, were rolled out in January of 2021. The various steps and methodology used for this analysis are stated below:

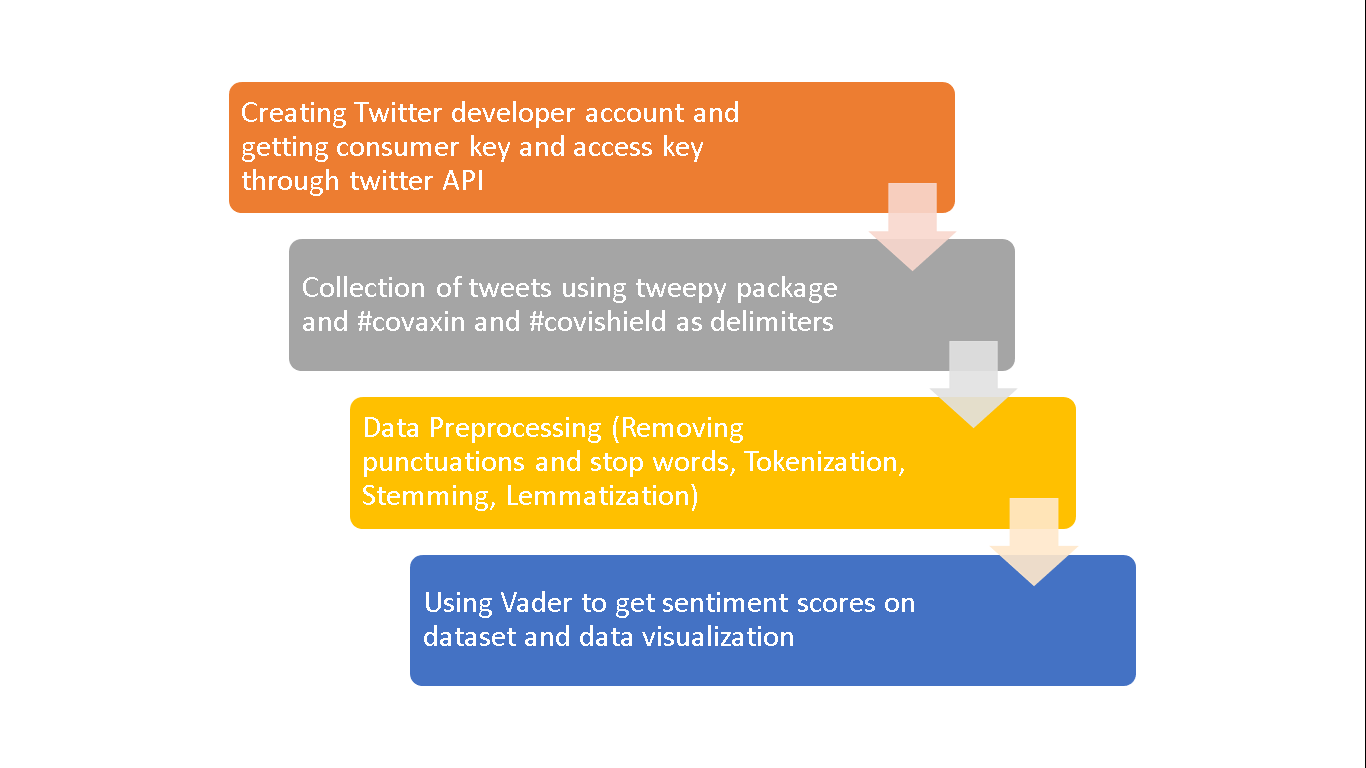


Fig. 3.4. Steps followed in Sentiment Analysis of tweets related to public reactions on vaccination drive

i. Creating Twitter developer account and generating necessary keys

The first step is related to data collection. For our data collection we needed tweets and the tweets can be taken from the twitter API using the twitter developer account. The twitter developer account was approved within days of applying. It is a platform which provides tools and resources and datasets for research and studies and finding solutions to real world problems. The twitter developer account is organised into 4 products namely – Labs, Twitter Ads API, Twitter for websites and Twitter API. For our purposes of analysis in our project, we have the Twitter API product. Twitter API is beneficial in finding and retrieving data from tweets, users, trends, media, direct messages, lists and places. It caters to people with varied interests with some people who wish to just use it for leaning purposes can get the standard package while qualified academic researchers may get the Academic Research package which will enable them to receive highly elevated access to endpoints.

We received the Academic Researcher package after applying for it. Our package had a limit of 500000 tweets which we can retrieve per month from the twitter database free of cost. Once we got the access. We created a project titled “COIVD-19\_senitmentanalysis” to generate consumer tokens and access tokens. The consumer tokens and access tokens are important keys which link our accounts to our codes and help us retrieve the tweets. These are unique keys and keeps a tab of the amount of tweets retrieved.

ii. Collection of tweets using Tweepy package

After the generation of the important keys, the next step was of data collection. The data was collected using the python code. Tweepy is a python library which was developed to access and use the twitter API using python. With the help of important keys of consumer token and access token, tweets were retrieved. The vaccines Covishield and Covaxin were developed in India and were rolled out by the Indian government on 16th January, 2021.

The three important factors for retrieving the tweets were the date, the hashtag and the number of tweets. For the hashtag we chose the delimiter #covaxin and #covishield because these two hashtags gave a better understanding of the tweets related to these two vaccines. For the date we decided on 4-5 days for both hashtags over a period of three months from January to March. We eventually decided 16th January to 19th January, 26th February to 28th February and 1st March to 5th March as dates for our analysis. The dates in January were taken keeping in mind the fact that the first phase vaccinations started on 16th January [9]. The dates in February were taken considering the events of our Prime Minister Mr. Narendra Modi taking the vaccine on 1st March[5]. All these dates were taken keeping in mind to see the effect of events and period on the sentiments of the masses in terms of positive, negative and neutral reactions. The third important factor was number of tweets which was decided at 1000 tweets per hashtag per day. The 1000 tweet limit was selected as it was a sizable sample size to do the analysis.

The data hence collected using the three factors was converted and stored in csv format. The csv format contained username, description, location, following, followers, total tweets, retweets, text and hashtags. The ‘text’ column contained the raw tweets which would eventually use as data for our analysis after cleaning it. Each csv was then merged to form a single csv containing data only for #covaxin. Same process was followed for #covisield csv and 3 csvs for the months January, February and March. Thus finally 5 different csv were prepared for analysis and the process of data collection was completed.

iii.Data Preprocessing

The third step of sentiment analysis was data preprocessing. It is a very critical step as the text based data used in our analysis contains a lot of unnecessary elements which has no weightage in determining the sentiment of a person and thus removing them will reduse the processing time of the model and visualizing the data. The data preprocessing of a twitter text based dataset was achieved through a multi stage process and we used the nltk package of python to do it. The nltk or natural language toolkit is used for working on data related to human language and facilitate text processing libraries which are crucial for preprocessing of the data. This multi stage process involves the following stages:

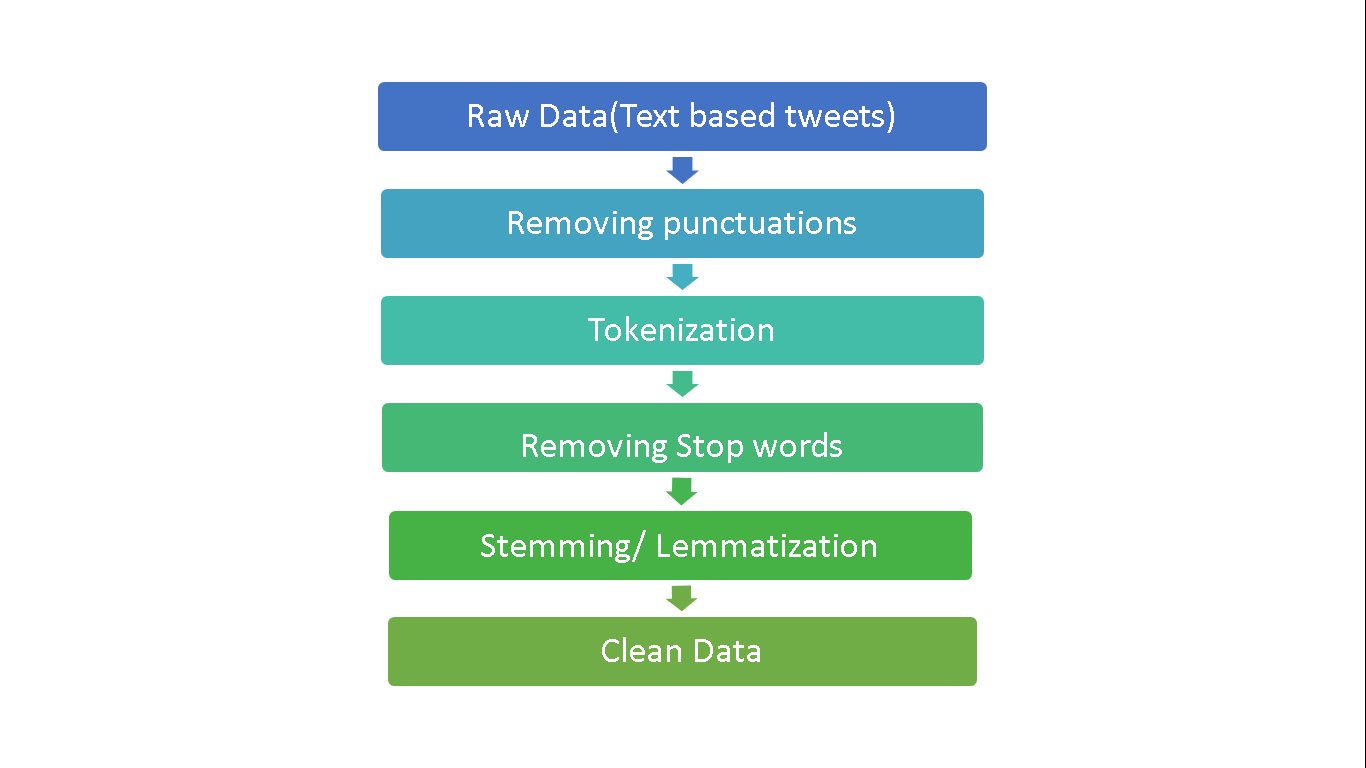


Fig. 3.5. Stages of Twitter Data Preprocessing

The raw data is the ‘text’ column in the dataset. For purposes of understanding the methodology at each step, we take the following tweet as an example:

"It’s not the to do thing, it’s a responsibility! Took 1st dose of #Covaxin 💉 at BKC MCGM Jumbo facility! I urge all individuals those are eligible for the vaccine please get it done! it’s safe! Please do wear Mask 😷 wash hands, keep urself sanitised &amp; maintain social distance! <https://t.co/qnh1ar7bO9>”

* 1. Removing Punctuations: The punctuations are of no use in sentiment analysis as they carry no information whatsoever related to sentiments of a person in English language so they are removed from the text dataset. The string package is used for this purpose and thus the “string.punctation” function is used for removing punctuations which contains the following punctuations- '!"#$%&\'()\*+,-./:;<=>?@[\\]^\_`{|}~'.

The data after removing punctuation looks like the following:

“It’s not the to do thing it’s a responsibility Took st dose of Covaxin 💉 at BKC MCGM Jumbo facility I urge all individuals those are eligible for the vaccine please get it done it’s safe Please do wear Mask 😷 wash hands keep urself sanitised amp maintain social distance httpstcoqnharbO”

* 1. Tokenization: The next step involves the splitting of sentences into words which is done using “re.split()” function . It splits token based on the white spaces and punctuations. The data after tokenization looks like the following:

“['it', 's', 'not', 'the', 'to', 'do', 'thing', 'it', 's', 'a', 'responsibility', 'took', 'st', 'dose', 'of', 'covaxin', 'at', 'bkc', 'mcgm', 'jumbo', 'facility', 'i', 'urge', 'all', 'individuals', 'those', 'are', 'eligible', 'for', 'the', 'vaccine', 'please', 'get', 'it', 'done', 'it', 's', 'safe', 'please', 'do', 'wear', 'mask', 'wash', 'hands', 'keep', 'urself', 'sanitised', 'amp', 'maintain', 'social', 'distance', 'httpstcoqnharbo']”

* 1. Removing Stop words: The search engines are programmed to ignore commonly used words in English language like ‘a’, ‘an’, ‘the’, ‘in’, ‘them’, etc. These words are known as stop words and hold no importance for sentiment analysis, hence they need to be removed during data cleaning. The function “ nltk.corpus.stopwords.words('english')” is used for this purpose. The data after removing stop words looks like the following:

“['thing', 'responsibility', 'took', 'st', 'dose', 'covaxin', 'bkc', 'mcgm', 'jumbo', 'facility', 'urge', 'individuals', 'eligible', 'vaccine', 'please', 'get', 'done', 'safe', 'please', 'wear', 'mask', 'wash', 'hands', 'keep', 'urself', 'sanitised', 'amp', 'maintain', 'social', 'distance', 'httpstcoqnharbo']”

1. Lemmatization and Stemming: The process of reducing inflected and derived words to their word stem or root form is stemming and the process of grouping together the inflected forms of a word so they can be analysed as a single item is called lemmatization. These are the last steps after which the data is clean and ready for analysis and visualization. The function “nltk.PorterStemmer()” is used for stemming and the function “nltk.WordNetLemmatizer()” is used for lemmatizing.

The stemmed data looks like the following:

“['thing', 'respons', 'took', 'st', 'dose', 'covaxin', 'bkc', 'mcgm', 'jumbo', 'facil', 'urg', 'individu', 'elig', 'vaccin', 'pleas', 'get', 'done', 'safe', 'pleas', 'wear', 'mask', 'wash', 'hand', 'keep', 'urself', 'sanitis', 'amp', 'maintain', 'social', 'distanc', 'httpstcoqnharbo']”

The lemmatized data which is the final pre-processed data looks like the following:

“['thing', 'responsibility', 'took', 'st', 'dose', 'covaxin', 'bkc', 'mcgm', 'jumbo', 'facility', 'urge', 'individual', 'eligible', 'vaccine', 'please', 'get', 'done', 'safe', 'please', 'wear', 'mask', 'wash', 'hand', 'keep', 'urself', 'sanitised', 'amp', 'maintain', 'social', 'distance', 'httpstcoqnharbo']”

iv. Sentitment Analysis using VADER and visualization of results

The final step of the project involves the use of VADER for labelling the tweets for different sentiments and visualizing the result. VADER stands for “Valence Aware Dictionary and sEntiment Reasoner” and it is a lexicon and rule-based sentiment analysis tool which is specifically costumed to sentiments which are conveyed and expressed on social media platforms. VADER uses a combination of sentiment lexicon is a list of lexical features such as words which are generally labelled according to how closely they tend to either positive or negative or neutral. VADER helps in dividing the data into positive, negative and neutral labels.

The function nltk.download(‘vader\_lexicon’) is used for downloading it. We get the polarity score for each tweet showing the positive, negative and neutral coefficient and the compound score using these three coefficients. If the compound score is greater than 0, then it is labelled as positive and if is less than 0, it is labelled as negative and neutral if 0.

The counting of positive, negative and neutral tweets in the dataset is then carried out. The result is visualised in the form of bar plot using the seaborn library. The matplotlib library is also used to generate the pie chart to understand how much percentage of these tweets possess positive sentiments, negative sentiments and neutral sentiments.

This whole process is followed for the five datasets. These 5 datasets include two datasets of #covaxin and #covishield collected combined for the months of January to March and the rest of the three datasets contains data for each month of January, February and March.

**CHAPTER 4**

**CONTRIBUTION OF EACH STUDENT**

*4.1 Contribution of Anmol Kumar*

i. Growth Analysis

For the growth analysis section of the project, I have contributed to the data collection process during the initial months of the project. This involved going through various datasets on the internet from government websites to university websites like that of John Hopkins University and private websites like ourworldindata.org and Kaggle. I found the “SARS 2003 OUTBREAK DATASET” on Kaggle and finalised it which has been used for the growth analysis section of the project. I selected this dataset based on the data fields which my partner and I identified during our weekly discussions. These fields were date, time, confirmed number of cases, confirmed number of deaths, confirmed number of recovered cases. The idea behind identifying these data fields was to make a model which shall predict how the cases will grow over a period of time. Another dataset – “COVID 19 INDIA” was collected by my partner.

After the datasets were collected, it was pre-processed by my partner. The pre-processed data was then used for growth analysis by testing it on three models namely linear regression and two non-linear models which were logistic regression and Gompertz model. I contributed to the linear regression modelling of the two data sets by writing the code for it. In the process of writing the code, I did a side project which involved linear regression to get a better grasp of the model. I trained and tested both the datasets by splitting randomly in a 30-70 split using the sklearn library to recognize trends and assess the model accuracy. I plotted the data using a scatter plot. It contains a line drawn which passes the scatter plot point and tries to fit maximum data points and thus this line is also called best fit line. The logistic regression modelling for both the datasets was done by my partner.

For the third model which was the Gompertz model, both of us tried to find and read any documentations and write the code. It was a combined effort which involved us brainstorming over zoom calls and debugging our code. Unlike the Linear Regression where the dataset was split randomly in a 30-70 format, in this case, trained the dates of the start of the time period and the testing was done on the latter dates. The training dataset recognizes patterns and the testing dataset is used to accurately predict the outcome. Hence, we use initial trends of the Covid-19 and SARS 2003 disease to predict how it will grow in the future. To plot the Gompertz curve, we chose a random state of India in the case of Covid-19 and a random country from the SARS dataset to showcase the growth curve. Lastly, we did curve fitting on the curve to provide a best fit to the specific dataset. The error and accuracy was then calculated by my partner for all the three models. Finally, we discussed the results of the three models and formed conclusions.

ii. Sentiment Analysis

After initial discussions with my partner and our guide, we decided to add the sentiment analysis section. We wanted analyse the views and opinions of people on twitter regarding vaccination drive. For the sentiment analysis section of the project, the data was collected in the csv format by my partner. I helped him by applying for twitter developer account since I had a twitter account which was 6 years old and thus it was easy to get approved quickly. 1000 tweets were scraped from the twitter API for each delimiter of #covaxin and #covishield for each day of 14th to 19th January and 26th February to 5th March and merged into five datasets. The five datasets were total tweets with Covishield and Covaxin and total tweets in the month of January, February and March. I wrote the code for pre-processing of these five datasets.

The pre-processing for text based twitter dataset is a multistage process. My partner and I divided different stages of the process and both of us did it. The nltk (Natural Language Toolkit) package is used for this process which provides good tools for cleaning text data. The steps for cleaning are removing punctuations, tokenization, removing stop words, stemming and lemmatization. I used the help nltk python documentations and articles of similar text based projects which needed pre-processing. The pre-processed data was also stored in csv format.

I worked on the final part of the sentiment analysis. The final step of this section of project involved the use of VADER for labelling the tweets for different sentiments and visualizing the result. VADER stands for “Valence Aware Dictionary and Sentiment Reasoner” and it is a lexicon and rule-based sentiment analysis tool which is specifically costumed to sentiments which are conveyed and expressed on social media platforms. VADER uses a combination of sentiment lexicon is a list of lexical features such as words which are generally labelled according to how closely they tend to either positive or negative or neutral. VADER helps in dividing the data into positive, negative and neutral labels. The function nltk.download(‘vader\_lexicon’) is used for downloading it. I generated the polarity score for each tweet showing the positive, negative and neutral coefficient and the compound score using these three coefficients. If the compound score is greater than 0, then it is labelled as positive and if is less than 0, it is labelled as negative and neutral if 0.

Following this, I plotted the bar plot using the seaborn library to visualise the count of positive, negative and neutral sentiments for each of the five datasets. Finally, I plotted the pie chart of percentage share of the three classifications of sentiments. The analysis of the results was done by both of us over zoom call and both of us equally contributed to the documentation and report preparation.

*4.2 Contribution of Tanmay Ambegaokar*

i. Growth Analysis

During the initial discussions with our internal guide, he mentioned doing a project on the Covid-19 pandemic among others. This idea stuck with us and we continued to find relevant information regarding the pandemic. The first month of this project, was spent largely in finding the correct datasets. Our criteria for finding the dataset was that it should be detailed, with a good number of columns. In our weekly calls, my partner and I finalized that the relevant fields must be date, time, State/Country, Number of confirmed cases, Confirmed Deaths, Number of Recovered Cases. We searched extensively on Data Science sites like Ourworldindata.org and Kaggle. We also searched in the official government sites of India. Finally, I found an extensive dataset on Kaggle on Covid-19 in India. It had all the fields which we had discussed along with some other fields. My partner, meanwhile, had found another dataset on SARS 2003 outbreak.

Once, both these datasets had been approved by our internal guide, I started with the next step of the project. The month of March was dedicated to pre-processing of data. This is one of the most important part as an unclean data will result in wrong predictions. It involves cleaning the data by removing unnecessary columns. The Covid-19 dataset had 2 extra fields related to confirmed Indian Nationals and Confirmed Foreign Nationals. We were not interested in both these fields for our analysis and hence was subsequently dropped using the .drop() function. Null values were also removed by replacing them with a 0 value.

The last two months of the project was spent on data visualization, modelling the data and evaluating the model. Over zoom calls both of us had decided to use three machine learning algorithms to model the data. Before modelling, I did the visualization part. It involved presenting the text data in the form of plots and graphs using matplotlib library of python. I also did the logistic regression algorithm on both the datasets. This is a non-linear regression model. This involved training and testing the datasets to get predictions on how the cases will grow. In the SARS dataset, we used confirmed cases as a metric to predict the growth over time. In the Covid-19 dataset, we used number of deaths as a metric. Linear Regression was done by my partner.

For the Gompertz model, we faced issues related to the date column. Both of us researched about the problem on Stack Overflow and was resolved by a combined effort. The final part of the growth analysis regarding the accuracy and error rate for all three models was done by me.

ii. Sentiment Analysis

The idea to do sentiment analysis had come up during the initial discussions with our internal guide. The largest vaccination drive in the world had just begun and there was a lot of debate regarding the efficacy of the vaccines because of a short trial phase. We decided it would be a good addition to the project to do a sentiment analysis using twitter tweets as our dataset to analyse the general reaction of the public with regards to both the vaccines being administered. I applied for a developer account on twitter which was approved in a week. Using this account, I could extract tweets in a csv format. The data was extracted using the hashtag “#covaxin” and “#covishield”. The csv file comprised of 1000 tweets from each day in the month of January, February and March. The csv file contained columns like name, number of followers, following, total tweets and the unclean tweet which I extracted.

Pre-Processing the data is the next step. Both of us divided the work and did the cleaning part using the nltk package. Punctuations and stop words were removed from the tweets and the clean data is again stored in the excel sheet.

The final part of the sentiment analysis was done by my partner using the VADER tool. Finally, over zoom calls, we completed the analysis part and the results were plotted on a

**CHAPTER 5**

**RESULT ANALYSIS**

* 1. *Growth analysis*

As mentioned earlier, we have two datasets on SARS 2003 epidemic and the Covid-19 in India dataset. We have done three kinds of regression on both these datasets.

1. Linear Regression

The data is plotted using a scatter plot and a line is drawn which passes the scatter plot point. It tries to fit maximum data points and thus this line is also called best fit line.

On the x-axis, we have the confirmed number of cases and on the y-axis we have taken the number of deaths recorded.

* SARS 2003 Dataset

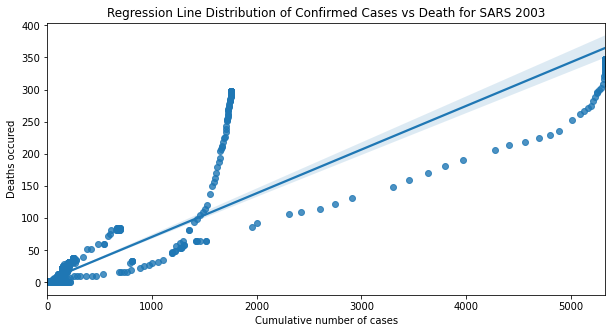


Fig 5.1. Regression Line Distribution of Confirmed cases vs Deaths for SARS 2003

To check the accuracy and error rate of the model, we have used two metrics to evaluate. Mean absolute error tells us the distance between the points and best fit line. If the data points sit perfectly on the line, the error rate will be 0. In our case, the MAE is 9.92. The co-efficient of regression needs to be closer to 1 for the model to be most accurate.

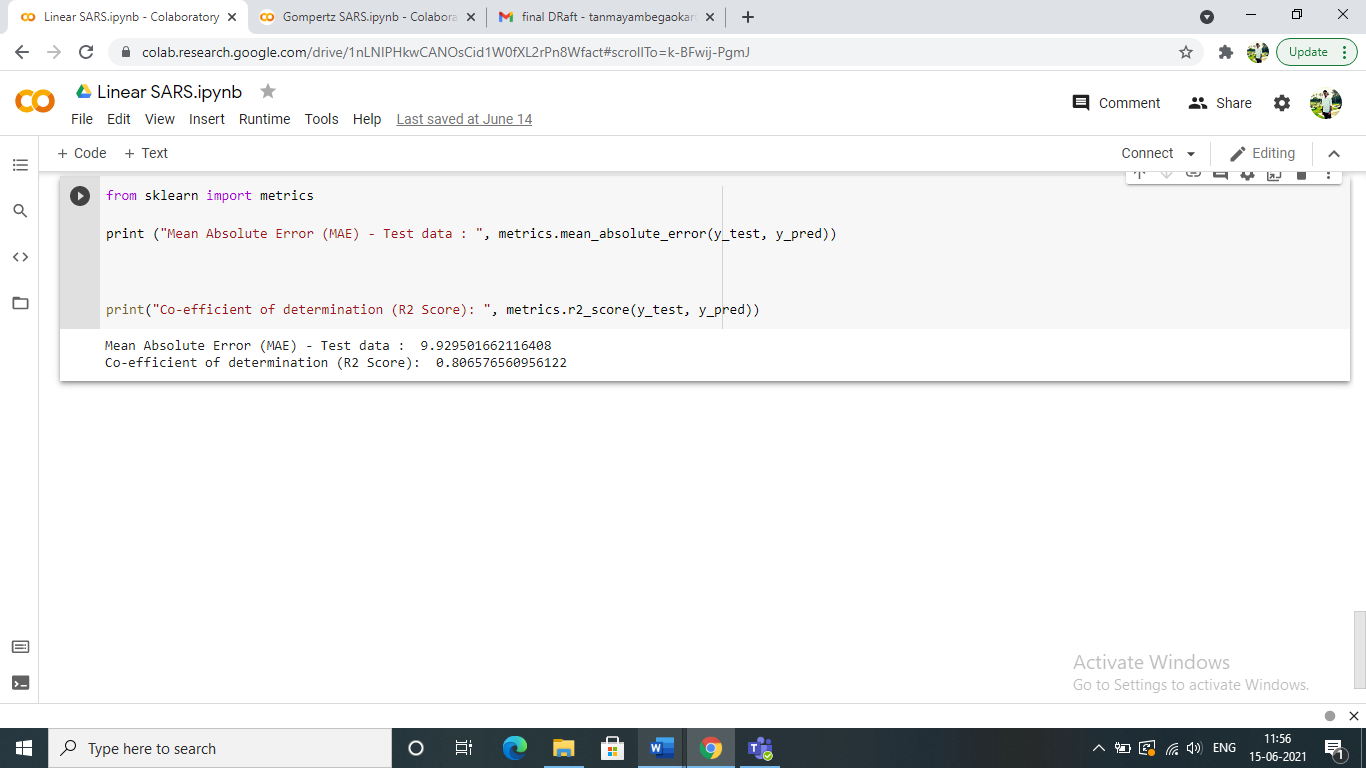


Fig 5.2 Error Rate and R2 score

* Covid-19 India Dataset

The same process has been followed as in the case of SARS dataset.

The mean absolute error for this model is 3.379 and the R2 score is 0.9012.

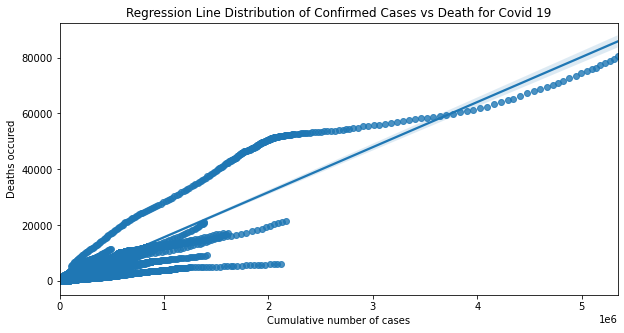


Fig 5.3 Regression Line Distribution of Confirmed cases vs Deaths for Covid-19(India)

1. Gompertz Curve

The Covid-19 cases and the SARS disease grows slowly at the beginning before flattening out in the end. In the middle, it follows a non-linear path. To plot the gompertz curve, we have chosen a random state of India in the case of Covid-19 and a random country from the SARS dataset to showcase the growth curve.

* SARS 2003 Dataset

For result purposes we have chosen Singapore as the target country to plot the graph. To plot any other country, we need to simply change the name of the country in the target country variable.

The left graph is the days vs confirmed cases graph. It shows how the cases grew with respect to the date. The plot on the right is the deaths vs days graphs. As mentioned earlier, it follows a non-linear path.

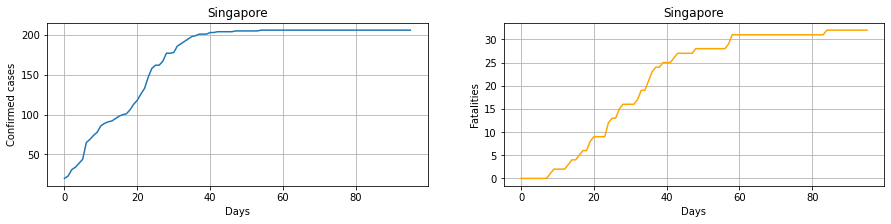


Fig 5.4 Gompertz curve for SARS 2003 in Singapore

The next step was to train the data and test it to predict how the cases will grow in the future. The orange line tells us how the cases are expected to increase in the future. The Mean Absolute Error for this model is 6.9526

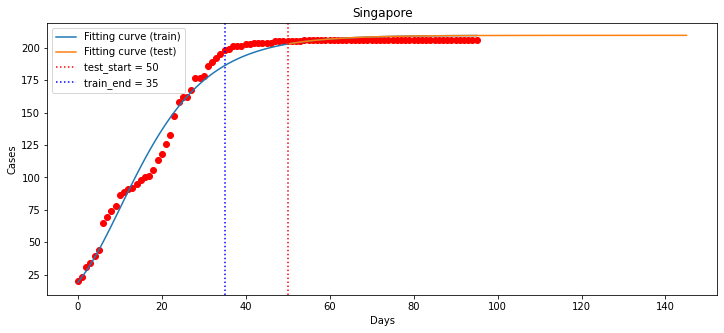


Fig 5.5 Curve fitting on Gompertz Curve- SARS 2003

* Covid-19 India Dataset

The same process is followed for this dataset also, the only difference being the country column is now replaced by the State column.

The left graph is the days vs confirmed cases graph. It shows how the cases grew with respect to the date. The plot on the right is the deaths vs days graphs.

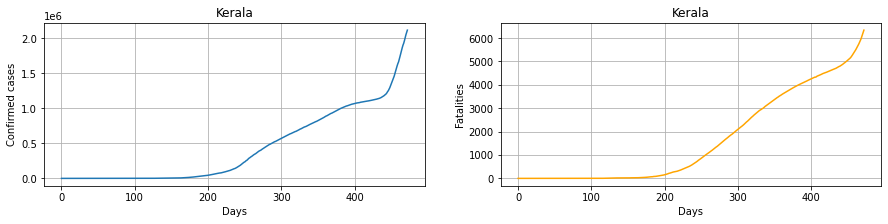


Fig 5.6 Gompertz curve for Covid-19 in Kerala

After the curve fitting step, we found out the MAE value to 20.123

The orange line shows us how the cases will increase 256 days after the first recorded case. Here, the lockdown restrictions are not taken into consideration and hence a slightly larger error rate.

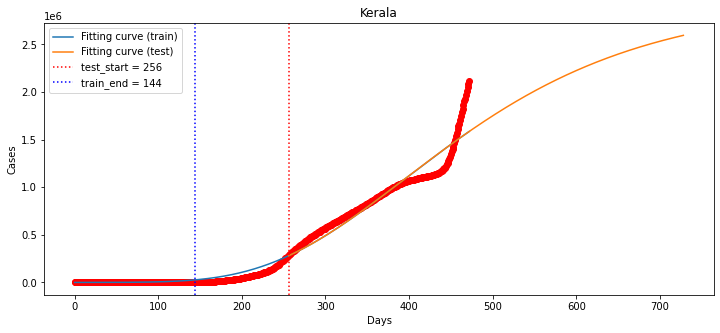


Fig 5.7 Curve fitting on Gompertz Curve- Covid-19(India)

1. Logistic Regression

Like the Gompertz curve, logistic curve is also a non-linear regression curve with a S-shaped curve. Th cases will grow slowly at the start and then grow exponentially till flattening out in the end.

* SARS 2003 Dataset

This plot tells the total confirmed cases of SARS 2003 and how the cases will grow and when it is expected to flatten out.

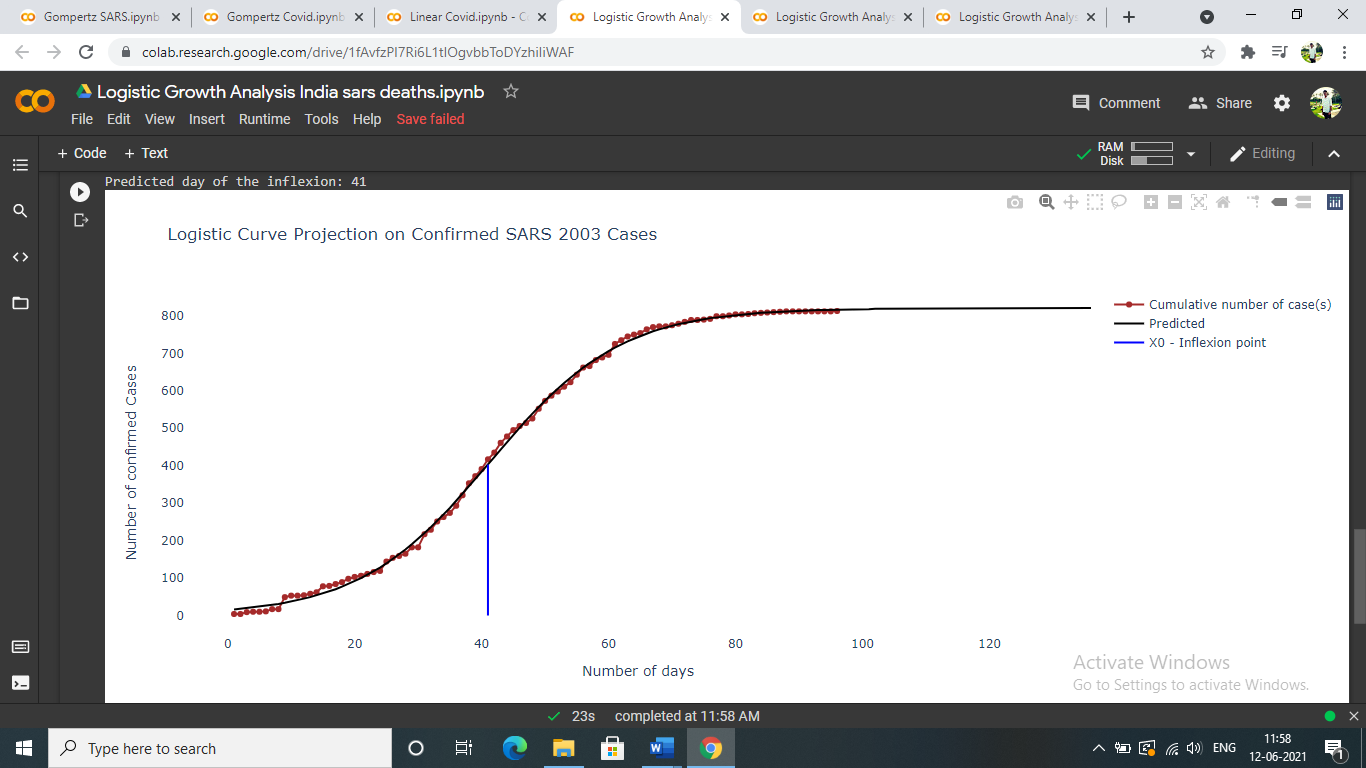


Fig 5.8 Logistic Regression Curve on SARS 2003

* Covid-19 India Dataset

This plot tells us the number of deaths vs the days since first death reported.

It also follows the non-linear path and flattens out in the end.

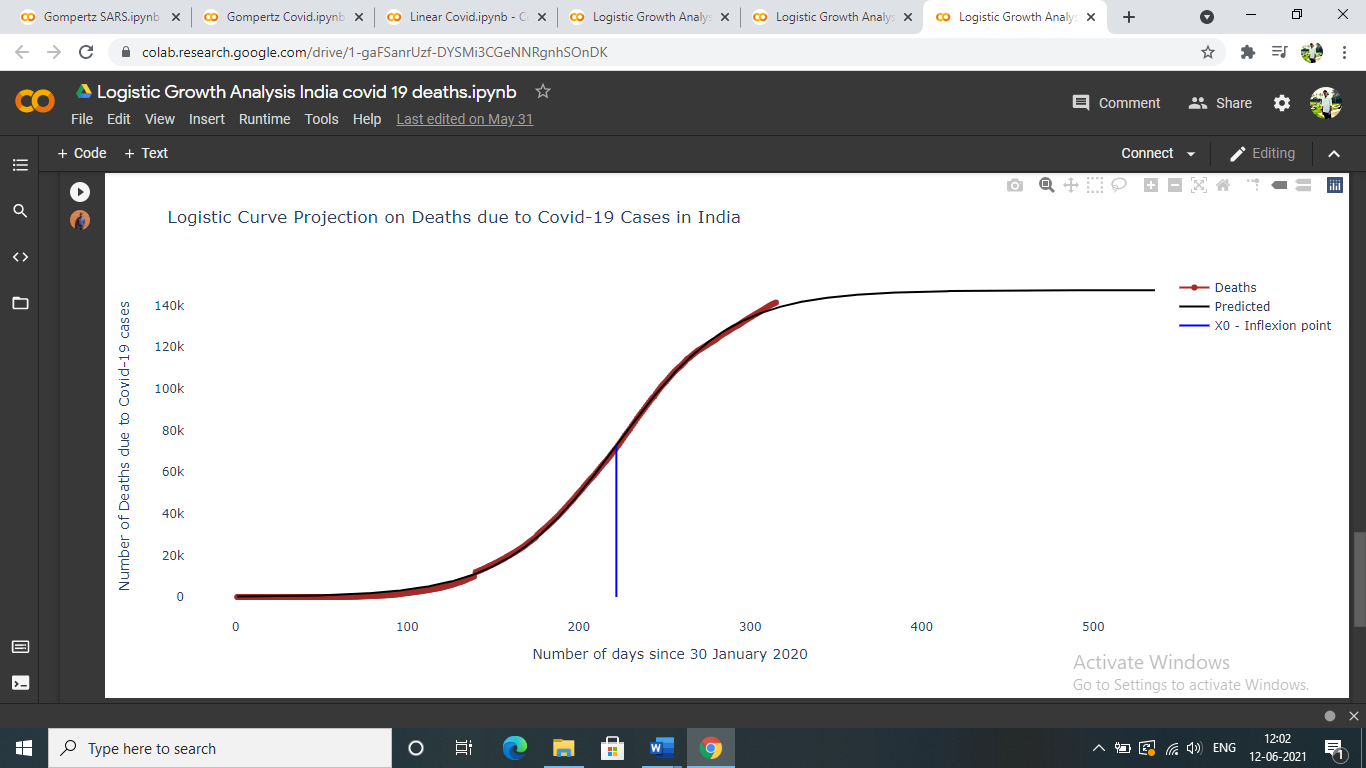


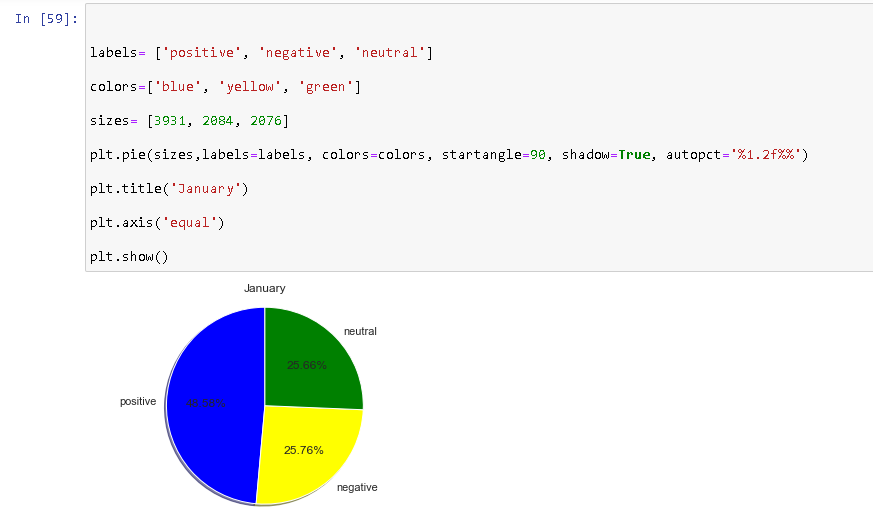
Fig 5.9 Logistic Regression Curve on Covid-19(India)

* 1. *The sentiment analysis on views and reactions for the vaccine of the Covid-19 virus*

The result of the sentiment analysis will be discussed in two different phases:

In the first phase of results, we look into month wise data and graphs of the positive, negative and neutral distribution of sentiments.

In the month of January, out of a sample size of 8091 tweets taken between 14th January and 19th January, 3931were positive, 2084 were negative and 2076 were neutral yielding a percentage share of 48.58%, 25.76% and 25.66% respectively.

 Fig 5.10 January Sentiments % pie chart Fig 5.11 January Sentiments Frequency Bar plot

In the month of February, out of a sample size of 5224 tweets taken between 26th February and 19th February, 2612 were positive, 1348 were negative and 1264 were neutral yielding a percentage share of 50.00%, 25.80% and 24.20% respectively.

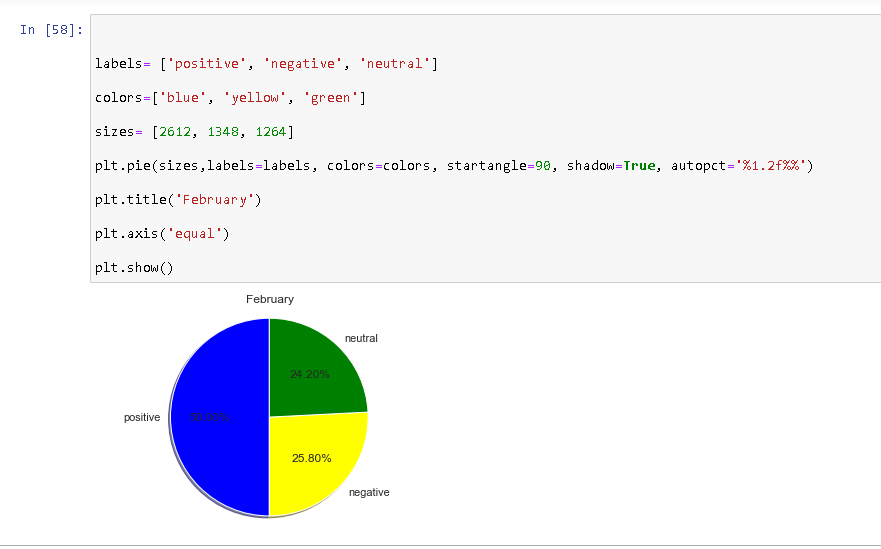
 

Fig 5.12 February Sentiments % pie chart Fig 5.13 February Sentiments Frequency Bar Plot

In the month of March, out of a sample size of 6741 tweets taken between 1st March and 5th March, 3251 were positive, 1785 were negative and 1705 were neutral yielding a percentage share of 48.23%, 26.48% and 25.29% respectively.

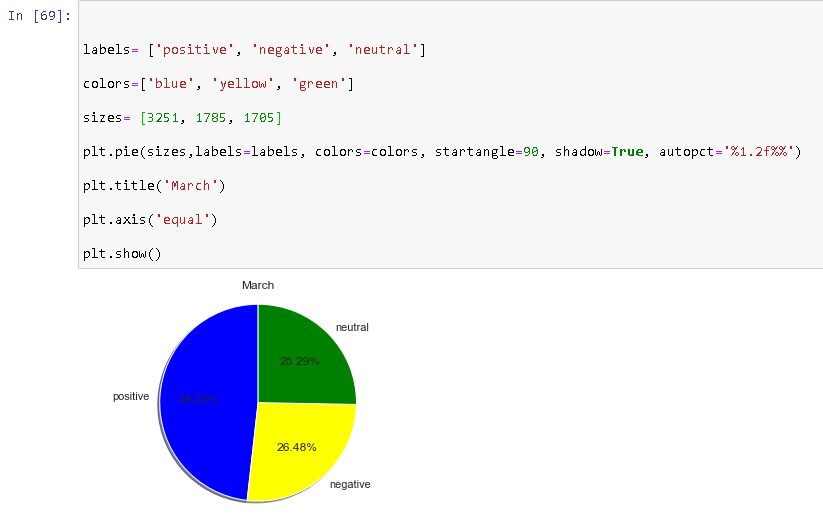
 

Fig 5.14 March Sentiments % pie chart Fig 5.15 March Sentiments Frequency Bar plot

In the second phase of results, we look into the data and graphs of the positive, negative and neutral distribution of sentiments for the vaccines Covaxin and Covishield in the months of January, February and March[4].

In the case of Covaxin, out of a sample size of 11176 tweets taken for the months of January, Februray and March, 5601 were positive, 3011 were negative and 2564 were neutral yielding a percentage share of 50.12%, 26.94% and 22.94% respectively.

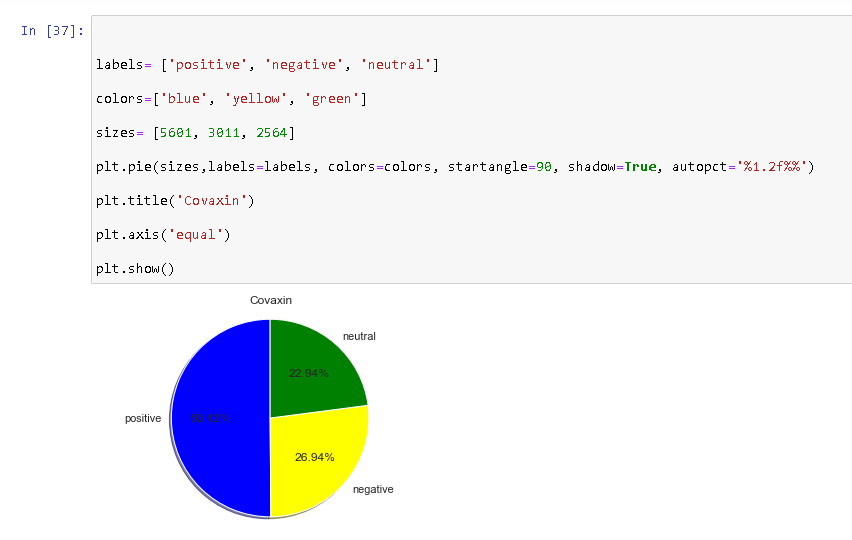
 

Fig 5.16 Covaxin Sentiments % pie chart Fig 5.17 Covaxin Sentiments Frequency Bar plot

In the case of Covicshield, out of a sample size of 8880 tweets taken for the months of January, Februray and March, 4193 were positive, 2653 were negative and 2034 were neutral yielding a percentage share of 47.22%, 29.88% and 22.91% respectively.

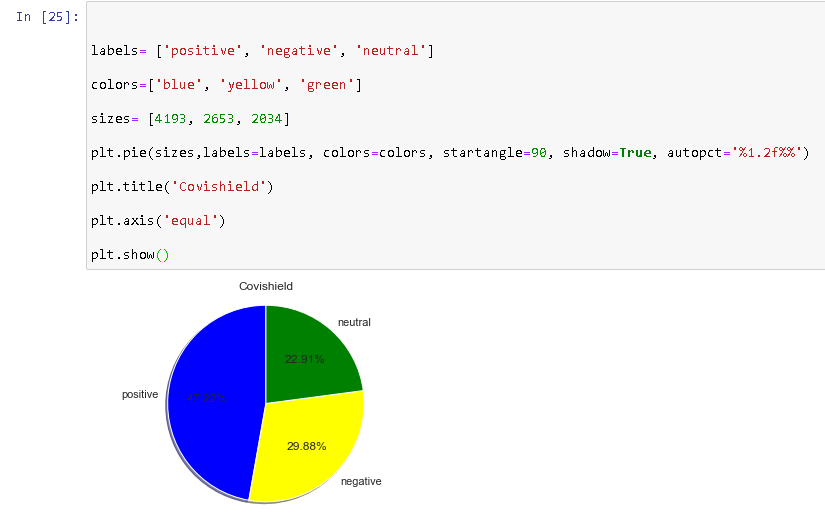
 

Fig 5.18 Covishield Sentiments % pie chart Fig 5.19Covishield Sentiments Frequency Bar plot

When we completed the sentiment analysis, we decided to do the same analysis for the month of June to see the current trend in opinions. The dataset had multiple same texts in tweets. When we analysed this, it presented a very skewed result towards positive percentage of tweets. This was because of a high number of duplicate tweets. When we tried to remove duplicate values, it made the dataset sample size too small to get a clear picture.

**CHAPTER 6**

**CONCLUSION AND FUTURE SCOPE**

* 1. *Growth Analysis*

Growth Analysis is done on the SARS 2003 and Covid-19 in India dataset using three machine learning algorithms. The algorithms used are linear regression, logistic regression, and Gompertz Curve model. These models were selected after thorough research before starting the analysis step. The two metrics chosen to evaluate the model were the mean absolute error and the co-efficient of regression value (R2 score).

The MAE gives the error rate of the model. If its closer to 0, it means the model is error free. Whereas, the R2 score gives the measure of the accuracy of the model. If the value is closer to 1, it tells us that the model is more accurate.

Table 5.3 Co-efficient of regression values for all models

|  |  |  |  |
| --- | --- | --- | --- |
|  | Linear Regression | Logistic Regression | Gompertz Curve |
| SARS 2003 DATASET | 0.8065 | 0.9219 | 0.8876 |
| Covid-19 India Dataset | 0.9012 | 0.951 | 0.7652 |

Table 5.4 Mean absolute Error for all models

|  |  |  |  |
| --- | --- | --- | --- |
|  | Linear Regression | Logistic Regression | Gompertz Curve |
| SARS 2003 DATASET | 9.925 | 1.236 | 6.9521 |
| Covid-19 India Dataset | 3.379 | 2.341 | 20.123 |

Looking at both these tables, we can see that the MAE score and the R2 score for the Logistic Regression model is closest to ideal model. Thus, we can conclude that among these three models, the Logistic Regression model is the most accurate and error free to predict the growth of coronavirus cases. We must bear in mind that the prediction will be true only if there is no change in external parameters.

* 1. Sentiment Analysis

The sentiment analysis on tweets related to vaccines Covishield and Covaxin gave us a fair understanding of how the general public felt when these vaccines were rolled out and were in circulation. The following table briefly summarises the results of the analysis.

Table 5.1 Sentiment analysis results for Vaccines

|  |  |  |  |
| --- | --- | --- | --- |
| Vaccine | Positive sentiment % | Negative sentiment % | Neutral sentiment % |
| Covaxin | 50.12 | 26.94 | 22.94 |
| Covishield | 47.22 | 29.88 | 22.91 |

Table 5.2 Month wise sentiment analysis results for both Vaccines

|  |  |  |  |
| --- | --- | --- | --- |
| Month | Positive sentiment % | Negative sentiment % | Neutral sentiment % |
| January | 48.58 | 25.76 | 25.66 |
| February | 50.00 | 25.80 | 24.20 |
| March | 48.23 | 26.48 | 25.29 |

The results in Table 5.1 above show that percentage of positive sentiments were more Covaxin as compared to Covishield. One possible conclusion for this may be because the efficacy rate of Covaxin is more than Covishield which might have factored into the opinions of the people. Secondly, a possible reason may also be that the period between two doses of vaccine is shorter for Covaxin than Covishield.

The results in Table 5.2 above show that the positive sentiments towards vaccines increased in the month of February but again decreased at the start of March. A possible explanation of this might be the start of second phase of vaccination for those who were of age 45 years and above and thus scepticism among them might be the reason for this.

The analysis classifies results in the form of positive, negative and neutral sentiments but the human emotions is much more complicated and varied than just these three. To further understand the sentiments on a deeper level, in future, the tweets could be classified into the human emotions of anger, happiness, sadness, trust, fear, joy and disgust. This might give a better explanation of the reactions of people towards the vaccination drive and the reasons behind those reactions.

**REFERENCES**

*Journal / Conference Papers*

1. V. Kharde and P. Sonawane, "Sentiment analysis of twitter data: A survey of techniques," arXiv preprint arXiv:1601.06971, 2016.
2. Modelling and Forecasting Covid-19 Growth Curve in India by Vikas Kumar Sharma
3. Tjørve KCM, Tjørve E. Modelling avian growth with the Unified-Richards: As exemplifed by wader-chick growth. *Journal of Avian Biology*. 2017.
4. Disinformation, ‘Fake News’ and influence Campaigns on Twitter by Matthew Hindman, George Washington University
5. Zarocostas, J., 2020. World Report How to fight an infodemic
6. Tausczik, Y.R., Pennebaker, J.W., 2010. The psychological meaning of words: Liwc and computerized text analysis methods. Journal of language and social psychology
7. Topic based Sentiment Analysis for COVID-19 Tweets Manal Abdulaziz1 , Mashail Alsolamy3 The Department of Information Systems King Abdulaziz University Jeddah, Saudi Arabia Alanoud Alotaibi2 The Department of Information Systems Imam Mohammad Ibn Saud Islamic University Riyadh,Saudi Arabia Abeer Alabbas4 Technical and Vocational Training Corporation Najran College of Technology Najran, Saudi Arabia
8. Bhat, M., Qadri, M., Noor-ul-Asrar Beg, M. K., Ahanger, N., & Agarwal, B. (2020). Sentiment analysis of social media response on the covid19 outbreak.
9. Public Sentiment Analysis of COVID19 Vaccination Drive in India by Dr. Akash D Dubey, Jaipuria Institute of Management, Jaipur

*Web*

1. <https://www.kaggle.com/imdevskp/sars-outbreak-2003-complete-dataset>
2. <https://www.kaggle.com/imdevskp/covid_19_india>
3. <https://science.thewire.in/health/india-covid-19-vaccination-drive-co-win-covaxin-covishield/>
4. https://indianexpress.com/article/india/pm-narendra-modi-coronavirus-vaccine-7208991/

**PROJECT DETAILS**

**Department of Electrical & Electronics Engineering**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Details of Student 1** | | | | | | | | |
| Name | Anmol Kumar | | | | | | | |
| Reg. No. | 170906482 | | Section | C | | Roll No. | | 3 |
| Mail ID | iqanmol@gmail.com | | | Mobile No | | 9113580610 | | |
| **Details of Student 2** | | | | | | | | |
| Name | Tanmay Ambegaokar | | | | | | | |
| Reg. No. | 170906292 | | Section | C | | Roll No. | | 36 |
| Mail ID | tanmayambegaokar@gmail.com | | | Mobile No  N | | 9819283430 | | |
| **Project Details** | | | | | | | | |
| Project Title |  | | | | | | | |
| Date of Reporting |  | Project Duration | | | | 4 months | | |
| **Guide details** | | | | | | | | |
| Name of Internal Guide | Prof. Satykam | | | | | | | |
| Designation, Department & Organization | Assistant Professor – Senior Scale, Electrical and Electronics Engineering | | | | | | | |
| Mail ID | satyakam.deo@manipal.edu | | | | Mobile | | 9663611063 | |