Sentiment Analysis of COVID-19

Project report submitted in partial fulfillment of the requirement for the degree

of

Bachelor of Technology

In

COMPUTER SCIENCE & ENGINEERING

` By

Rahul Rana (161214)

Under the supervision of

Dr. Mrityunjay Singh

to

Department of Computer Science & Engineering and Information Technology

Jaypee University of Information Technology, Waknaghat, Solan-173234,

Himachal Pradesh

Certificate

I hereby declare that the work presented in this report entitled **"Sentiment Analysis of COVID-19 "** in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science & Engineering** submitted in the Department of Computer Science & Engineering and Information Technology**,** Jaypee University of Information Technology, Waknaghat is an authentic record of our own work carried out over a period from May 2020 to June 2020 under the supervision of **Dr. Mrityunjay Singh.**

The matter embodied in the report has not been appeased for the award of any other degree or diploma.

Rahul Rana

161214

This is to certify that the above affirmation made by the candidate is true to the best of my knowledge.

……………………………

Dr. Mrityunjay Singh
Assistant Professor (SG)
Department of Computer Science & Engineering and Information Technology
Jaypee University of Information Technology
Dated: June 23, 2020

Acknowledgement

I would like to take the opportunity to thank and express my deep sense of gratitude to my mentor and project guide Dr. Mrityunjay Singh for his immense support and valuable guidance without which it would not have been possible to reach this stage of my final year project.

I am also obliged to all my faculty members for their valuable support in their respective fields which helped me in reaching at this stage of my project. My thanks and appreciations also go to the colleagues who have helped me out with their abilities in developing the project.

Rahul Rana

161214

Date: June 23, 2020

Table of Contents

List of Figures

ABBREVATIONS

ML	: Machine Learning
BERT`	: Bi-directional encoding representation for transformer
IDE	: Integrated Development Environment
API	: Application Programming Interface

ABSTRACT

Corona virus or covid-19 is virus which started in the Wuhan province of China. After its outbreak in various parts of the world people got really curios about what exactly this is and how will it affect human life. This project aims in analyzing the sentiments of the people during this pandemic. Firstly reading about the existing work and getting to know about the technologies to get a overview of the project. Data was then collected from online source like Twitter using the scrapper provided. This dataset had different records for each date in English language. Later pre processed and few features were selected accordingly. This final dataset was filtered to produce a second dataset which had tweets from India only. Both the data frames were used to show insights about various measures like mean likes, retweets, polarity value of the polarity and subjectivity over the period, wordcloud showing the different types of words used in the text. Finally using the bert model for emotion classification. These experiments helped in getting to analyze the sentiments of the people.

CHAPTER 1

INTRODUCTION

CHAPTER 1: INTRODUCTION

1.1 Introduction

Nowadays, everyone is curious about what people think about a particular product, organization or a situation. These can be predicted from what they speak or what they write about it. Here comes the concept of sentiment analysis.

Sentimental analysis is where you interpret the sentiments or emotions of the people and classify them into various types, keeping in order various factors. This helps any organization in knowing the people's emotions and act accordingly. Sentiment analysis is divided into various types depending on what outcome one is looking for. Analysis can be made on the polarity of the text which ranges from very negative to very positive and even neutral; it can be based on the emotions of the text; feedback about a particular feature i.e. aspect based and finally analysis of text involving different languages, where you have to detect the language as well.

There comes the question that what is the need of analyzing the sentiments. This analysis uses large amount of data which is initially not structured. All the data is extracted and made clean with preprocessing, which helps in creating maintainable datasets. Real time analysis helps the organization look at the current scenario and make future decisions accordingly and gain better results.

1.2 Problem Statement

Corona virus or covid-19 is virus which started in the Wuhan province of China. After its outbreak in various parts of the world people got really curios about what exactly this is and how will it affect human life. There were lockdowns put into action around the world and people had different views about this situation. People currently in different states and countries wanted to go back home safely, flights were sent to bring the citizens back. Every country with increasing cases had to face fatal cases and also the decline in the economy of the country. Countries suffered from this outbreak could not find the vaccination for the virus and the constant increase is still a major problem in many countries. People came together on social networks showing helping hand in the form of donations to the country to fight back the virus. China on the other hand was targeted for the cause

We know that social media platforms are the best places to vent out or discuss any ongoing issues. Hence, it is the best source to view their emotions. I will be performing sentiment analysis on the same and checking what are the different opinions that came out for this pandemic.

1.3 Objectives

The main objective of the project would be analyzing tweets from twitter to draw insights about the data. There will be certain steps involved to carry out this task;

a) Collecting tweets from twitter
b) Cleaning the tweets.
c) Viewing relevant insights from the data.
d) Classifying the data

After performing these steps, I will come to a conclusion on how the outbreak affected people in India.

1.4 Methodology

I will be scraping the tweets from twitter using the scrapper API and the tweepy library provided by python. These will be filtered according to the requirement and to make it country specific. Later these tweets will be cleaned so as to get clean data which will be easy to use and will get better insights. The dataset will have various features, for which analysis can be performed but only some of the features will be selected to carry out the further process.

Insights will be drawn from the final cleaned data. These insights will require various measures like polarity, subjectivity and wordcloud showing the different types of words used in the tweets. Finally Bert model will be used for emotion classification which will be discussed later in the report.

1.4.1 Performing sentiment analysis

Natural Language Processing provides various algorithms to perform analysis on textual data. These algorithms work on different parameters which ranges from manual work to automatic process by in built libraries.

Rule based approach is the one where we manually define rules to perform stemming and covering into tokens and then defining set of words as positive, neutral and negative. These sets are then used to calculate polarity of the sentences where the maximum from a particular set decides the polarity of the text. Another approach is the automatic where all the above mentioned process is done by machine learning techniques and you get the desired result. Fig. 1.1 shows the working of a sentiment analysis model. You can also use a mix of these two techniques to make your model.

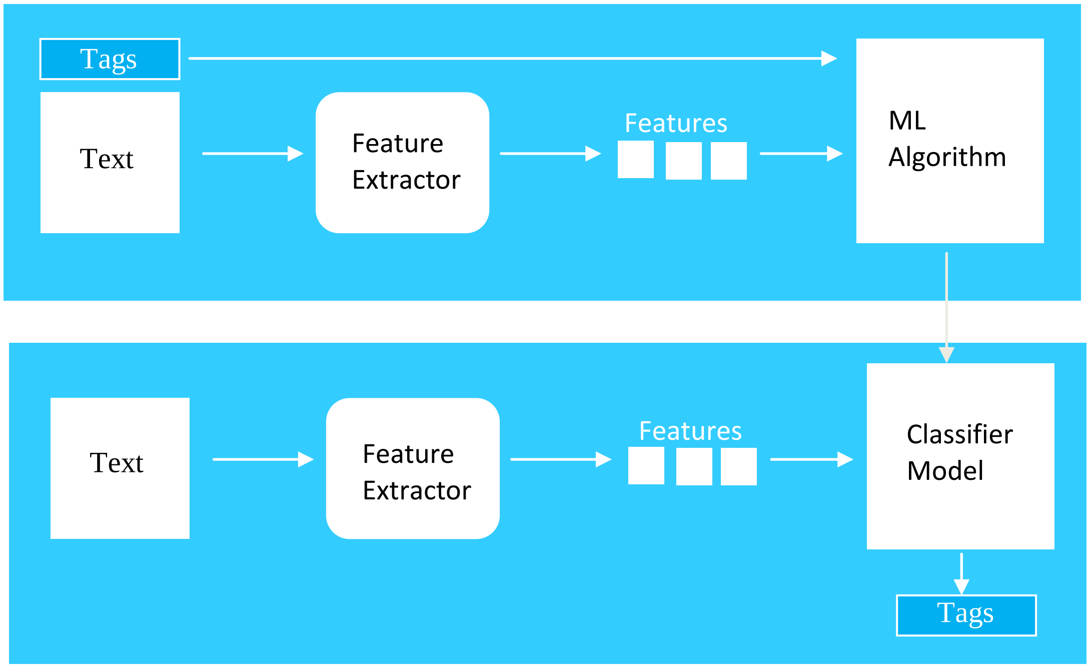


Fig 1.1 Working of sentiment analysis

The figure shows how the process is divided into two phases; training and prediction. The training phase includes the tagging of the data, feature extraction and using a machine learning algorithm to make a model to be used in the prediction phase, where more textual data is used to predict tags accordingly.

1.5 Organization of the thesis

The presented thesis work is divided into 5 chapters. In the first chapter, the problem is stated about why I am carrying out this analysis, the objectives have been formulated and the methodologies that will be used are defined. The second chapter will show the literature review carried out and the existing methods. In the third chapter, the design of experiment is discussed and what will be the system requirements. The fourth chapter will have the performance analysis and the results of the analysis and the final chapter will share the conclusion of the work carried out and future work if any.

CHAPTER 2

LITERATURE SURVEY

CHAPTER 2: LITERATURE SURVEY

2.1 Introduction

Corona virus, with its first case in Wuhan, China started to spread in various parts of the country and world. People returning from China were also getting affected by the virus. People used social media platforms to express their emotions and this pandemic brought out various emotions which were positive, negative as well as unbiased and showed emotions like fear, support etc.Since this a new topic people are still working on it and trying to analyze the scenario. Many researchers have worked on predicting things like by what time certain country will see a relief from the virus and researches based on predicting the number of confirmed cases based on the trend of curves plotted on the current situation and how it increased or decreased from past. There are many research papers that show how and when sentiment analysis can be used. What are the benefits, what should be the best source for data? This chapter will focus on looking for previous work in the relevant field and gaining knowledge about the same.

2.2 Need for sentiment analysis

The term sentiment analysis is also term as 'opinion mining' [3] which makes uses of natural language processing techniques to analyze a person's opinion, emotion and passionate response to situations, products or about another person. Showing this kind of information and presenting it in a graphical manner has gained much importance in the recent times and is a attractive research topic. [2]

2.3 Previous work

I went through various works that showed how people have used various methodologies to perform sentiment analysis. This involved extracting lexical sentiments with the documents [6], using bi an tri-grams [4] to see how features are associated with the sentiment, emotions are now a common way to express feelings hence emojis for positive, neutral and negative words [5]. Fig 2.1 shows the use of one of the methods to perform sentiment analysis and the system works internally. How data is read from the documents and can be subdivided into various sentiment groups like positive, negative and neutral. Al

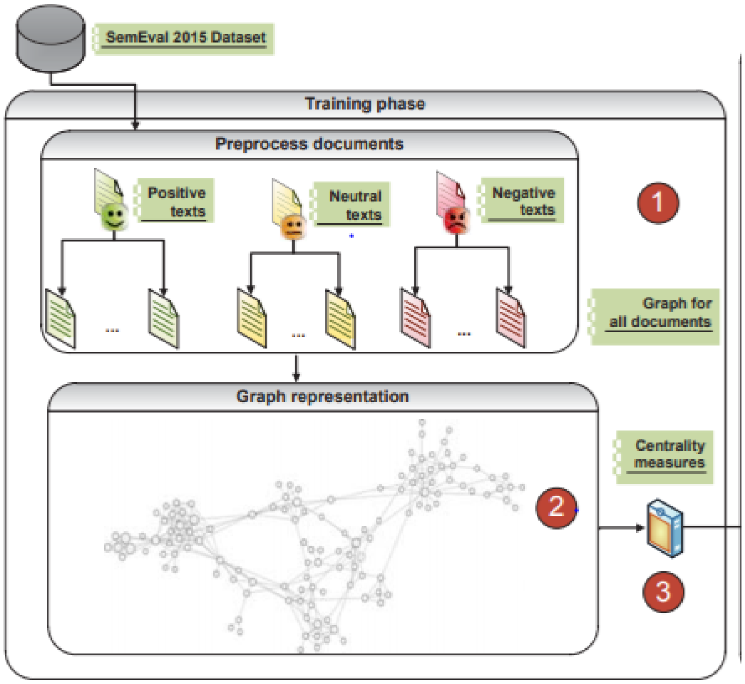
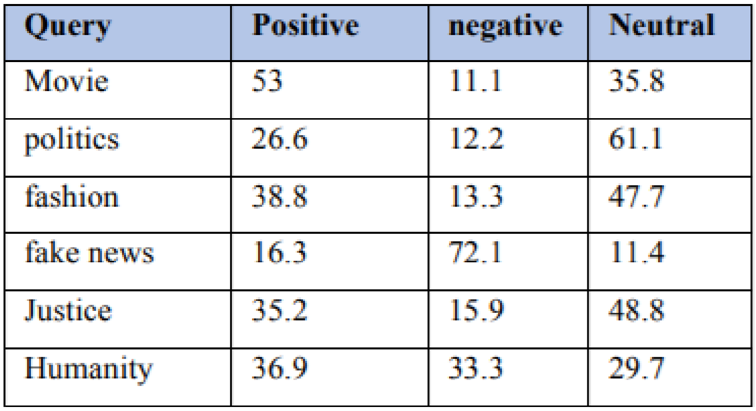


Fig 2.1 Shows representing results into graphical format

These works show what operations can be performed on the data to analyze it and bring out results like opinions, emotions and feedback as mentioned earlier. As seen from the research people often used graph based representation methods to show the output. This is mainly because it looks more attractive and people can easily judge on the facts from it.

2.4 Social Networks

A question comes up that why social networks are the best platform to get desired data for any scenario. The numbers of users have increased over the time and you can get large data from these sites like facebook, YouTube and twitter [8]. People spend hours on these sites [9] looking for various things like new posts, what's going on around them, what are their friends upto etc. Hence this makes these sites the most valuable source of data which can be used for various tasks. Even these sites cannot save all the data, since its very large when a time of years is considered. Big data is hence used to provide us access to the same using API calls. Fig 2.2 shows the data collected for some keywords



Query	Positive	negative	Neutral
Movie	53	11.1	35.8
politics	26.6	12.2	61.1
fashion	38.8	13.3	47.7
fake news	16.3	72.1	11.4
Justice	35.2	15.9	48.8
Humanity	36.9	33.3	29.7

Fig 2.2 the dataset with query keywords and the resultant percentage of each sentiment group

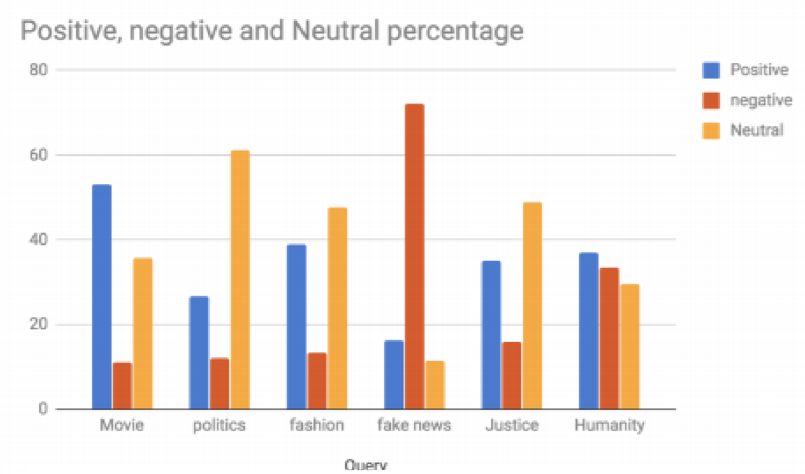


Fig 2.3 Shows the graphical representation of the above results

used to extract data from twitter [9] and columns showing the percent of each of the 3 categories calculated. Then the graphical representation about the same helps in viewing the results in a more exiting way. These results help any organization in knowing the brand value id the tweets are filtered according to the need and specific keywords. Later it will help them stand out as they will know what went wrong and where the need to improve themselves.

Why Twitter?

Twitter has a wide variety of people ranging from a common man to the presidents of the country. Even celebs, sports personalities and the top leaders from various organizations can [10] be seen on twitter expressing themselves about day to day activities and the current situation. Therefore it becomes the most valued platform for data collection as it will provide large amount of data plus data from different types of lifestyles.

2.5 BERT

BERT is a model for language processing built on bi-directional training and transformer, which are several model variants based on model configurations [11]. People have used BERT- base as their base model; which had 12-layer transformer blocks and for each block 12-head self attention layer and 768-dimensional layer. This takes in input 512 tokens at max and outputs the vector representations. Fig 2.4 shows the visual of the

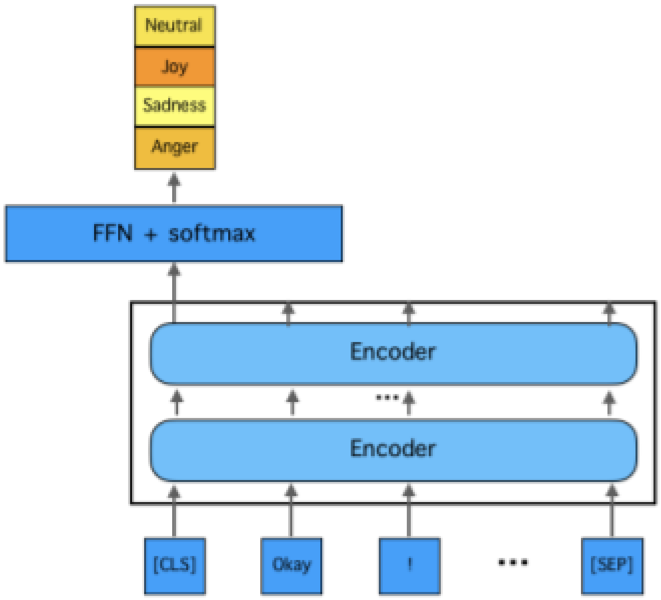


Fig 2.4 Working of the BERT text classifier using encoders

Internal working of the pre-trained BERT model [11]. Where the output can be seen as the emotions detected from the data which is labelled in the training phase and the testing phase predicts the output for other textual data.

2.6 Conclusion

After studying the existing work, I came to know about various technologies which can be used to perform sentiment analysis and how graphical representations can be an effective way to show the analysis performed. Learned about why social media platform like Twitter is best suited for data scraping. Also the use of Bert model for text classification. These techniques will be used accordingly.

CHAPTER 3

SYSTEM DEVELOPMENT

CHAPTER 3: SYSTEM DEVELOPMENT

3.1 Introduction

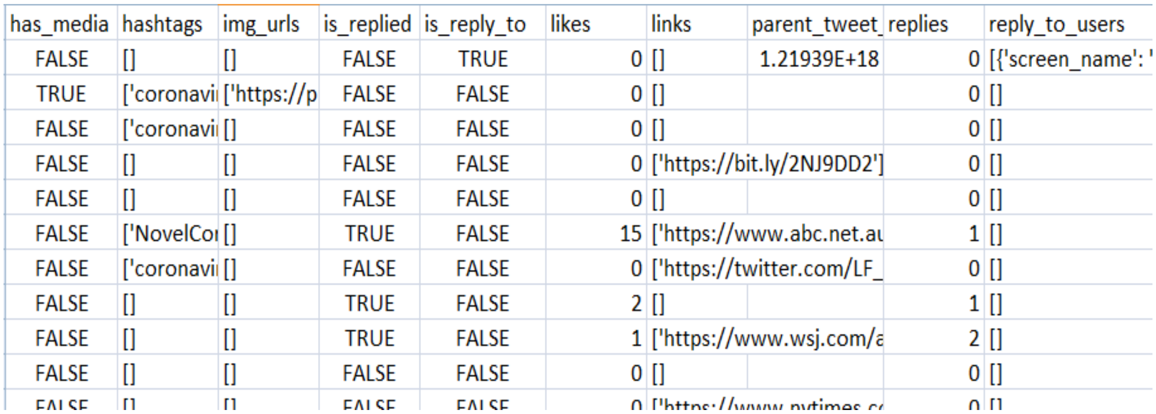
In this chapter steps for developing a system will be discussed. After learning about the process of sentiment analysis and the methods people have used to perform similar operations, I will build my project accordingly, so as to get better results. This will also aim on making a clean and understandable dataset as it is the main component of any machine learning algorithm.

3.2 Corona Virus

As discussed earlier this virus had a major outbreak around the various parts of the world and people were affected at a large scale. This lead to major loss in human life. People had different views about the same which is the aim of this project i.e. to carry out analysis on the views of the people, what went wrong for them.

3.2.1 Dataset

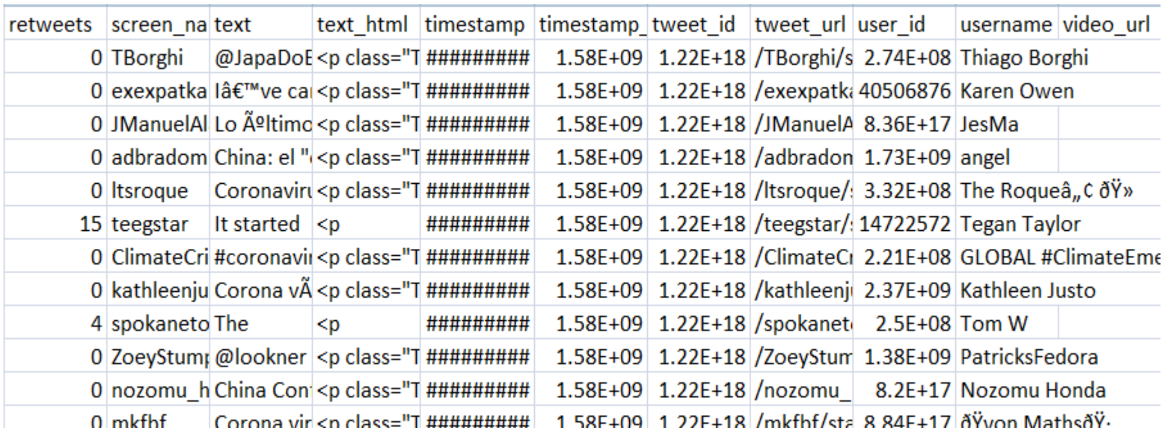
Twitter was chosen as the social media platform for extracting views in the relevant topic. This data was scrapped using the twitters scrapper API and the twitter search API, keeping in mind the restrictions of data use. These APIs provide data which has various features about a particular tweet



has_media	hashtags	img_urls	is_replied	is_reply_to	likes	links	parent_tweet_	replies	reply_to_users
FALSE	[]	[]	FALSE	TRUE	0	[]	1.21939E+18	0	[{'screen_name': '
TRUE	['coronavi	['https://p	FALSE	FALSE	0	[]		0	[]
FALSE	['coronavi	[]	FALSE	FALSE	0	[]		0	[]
FALSE	[]	[]	FALSE	FALSE	0	['https://bit.ly/2NJ9DD2']		0	[]
FALSE	[]	[]	FALSE	FALSE	0	[]		0	[]
FALSE	['NovelCo	[]	TRUE	FALSE	15	['https://www.abc.net.au		1	[]
FALSE	['coronavi	[]	FALSE	FALSE	0	['https://twitter.com/LF_		0	[]
FALSE	[]	[]	TRUE	FALSE	2	[]		1	[]
FALSE	[]	[]	TRUE	FALSE	1	['https://www.wsj.com/a		2	[]
FALSE	[]	[]	FALSE	FALSE	0	[]		0	[]
FALSE	[]	[]	FALSE	FALSE	0	['https://www.nytimes.c		0	[]

Fig 3.1 Data from scrapper I

Fig 3.1 and 3.2 show the different attributes that were derived from the tweets. Using these attributes I will perform operations.



retweets	screen_na	text	text_html	timestamp	timestamp_	tweet_id	tweet_url	user_id	username	video_url
0	TBorghi	@JapaDoE	<p class="T	#########	1.58E+09	1.22E+18	/TBorghi/s	2.74E+08	Thiago Borghi	
0	exexpatka	Iâ€™ve ca	<p class="T	#########	1.58E+09	1.22E+18	/exexpatk	40506876	Karen Owen	
0	JManuelAl	Lo Ãºltimo	<p class="T	#########	1.58E+09	1.22E+18	/JManuelA	8.36E+17	JesMa	
0	adbradom	China: el "	<p class="T	#########	1.58E+09	1.22E+18	/adbradon	1.73E+09	angel	
0	ltsroque	Coronaviru	<p class="T	#########	1.58E+09	1.22E+18	/ltsroque/	3.32E+08	The Roqueâ„¢ ðŸ»	
15	teegstar	It started	<p	#########	1.58E+09	1.22E+18	/teegstar/	14722572	Tegan Taylor	
0	ClimateCri	#coronavi	<p class="T	#########	1.58E+09	1.22E+18	/ClimateC	2.21E+08	GLOBAL #ClimateEme	
0	kathleenju	Corona vÃ	<p class="T	#########	1.58E+09	1.22E+18	/kathleenj	2.37E+09	Kathleen Justo	
4	spokaneto	The	<p	#########	1.58E+09	1.22E+18	/spokanet	2.5E+08	Tom W	
0	ZoeyStump	@lookner	<p class="T	#########	1.58E+09	1.22E+18	/ZoeyStum	1.38E+09	PatricksFedora	
0	nozomu_h	China Con	<p class="T	#########	1.58E+09	1.22E+18	/nozomu_	8.2E+17	Nozomu Honda	
0	mkfbf	Corona vir	<p class="T	#########	1.58E+09	1.22E+18	/mkfbf/sta	8.84E+17	ðŸvon MathsðŸ·	

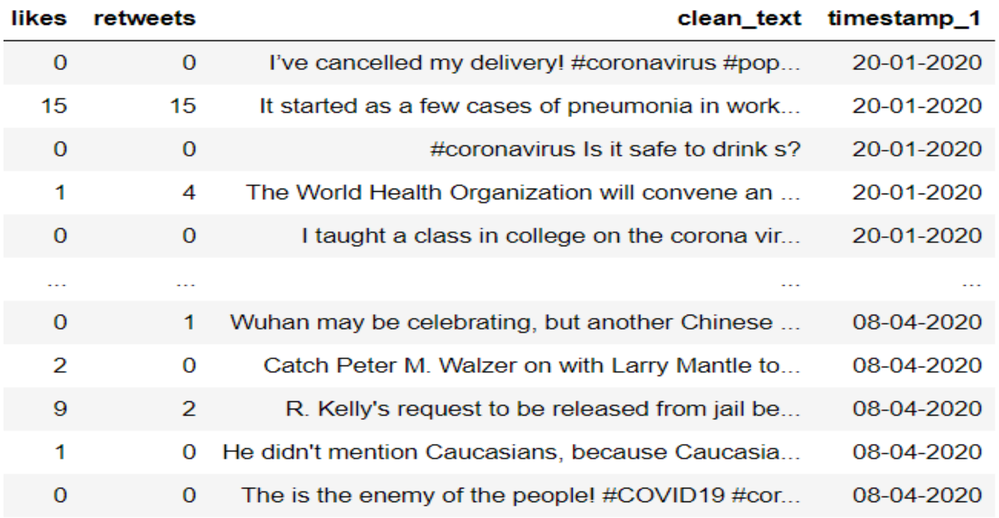
Fig 3.2 Data from scrapper II

3.2.2 Feature Selection

The dataset had many parameters like has_media, hashtags, img_urls, is_replied, is_reply_to, likes,links, parent_tweets, replies, reply_to_users, retweets, screen_name, text_html,timestamp,tweet_id,tweet_url,user_id,username and video_url. But only a few of them were required. The dataset created had only id, likes, and retweets, timestamp and the text of the tweet.

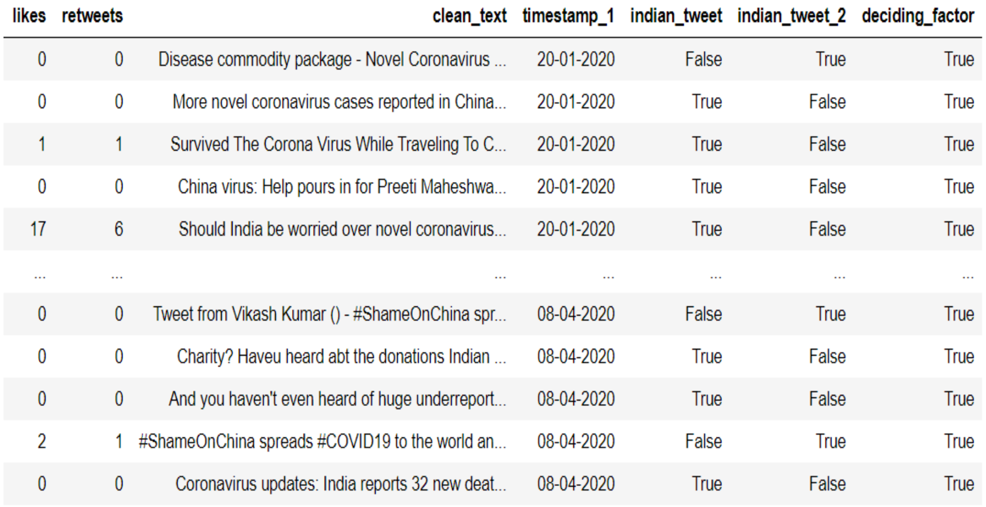
This tweet text needs to be cleaned before starting the analysis process. Cleaning of tweets is done using regression where tweet text is mapped to a equation which filters out links, images and emoticons from the text. Time stamps have date and time of the tweet but only date is required. Twitter scrapper API is used to extract data for each date from 20 Jan 2020 to 25 April 2020 using the begin and end time for each statement with words hashtags like: (#COVID2019 OR #COVID19 OR corona&virus) to get the results. Two data frames were created, one from the entire world and one for analyzing India specific data. For this purpose keywords like (India and Modi) were used to filter out tweets from India. The resultant dataset had columns defining for which measure it has been selected; i.e. Indian tweets combined using Indian tweets 1 and Indian tweets 2. If either one of this

Is available in the text, it is selected for the analysis. I couldn’t find appropriate method to use location filter so specific keywords could only be used to get the second data frame.



likes	retweets	clean_text	timestamp_1
0	0	I’ve cancelled my delivery! #coronavirus #pop...	20-01-2020
15	15	It started as a few cases of pneumonia in work...	20-01-2020
0	0	#coronavirus Is it safe to drink s?	20-01-2020
1	4	The World Health Organization will convene an ...	20-01-2020
0	0	I taught a class in college on the corona vir...	20-01-2020
...	...	...	...
0	1	Wuhan may be celebrating, but another Chinese ...	08-04-2020
2	0	Catch Peter M. Walzer on with Larry Mantle to...	08-04-2020
9	2	R. Kelly's request to be released from jail be...	08-04-2020
1	0	He didn't mention Caucasians, because Caucasia...	08-04-2020
0	0	The is the enemy of the people! #COVID19 #cor...	08-04-2020

Fig 3.3 Final dataset created after cleaning (World dataframe)



likes	retweets	clean_text	timestamp_1	indian_tweet	indian_tweet_2	deciding_factor
0	0	Disease commodity package - Novel Coronavirus ...	20-01-2020	False	True	True
0	0	More novel coronavirus cases reported in China...	20-01-2020	True	False	True
1	1	Survived The Corona Virus While Traveling To C...	20-01-2020	True	False	True
0	0	China virus: Help pours in for Preeti Maheshwa...	20-01-2020	True	False	True
17	6	Should India be worried over novel coronavirus...	20-01-2020	True	False	True
...	...	...	...	...	...	...
0	0	Tweet from Vikash Kumar () - #ShameOnChina spr...	08-04-2020	False	True	True
0	0	Charity? Haveu heard abt the donations Indian ...	08-04-2020	True	False	True
0	0	And you haven't even heard of huge underreport...	08-04-2020	True	False	True
2	1	#ShameOnChina spreads #COVID19 to the world an...	08-04-2020	False	True	True
0	0	Coronavirus updates: India reports 32 new deat...	08-04-2020	True	False	True

Fig 3.4 Final dataset created after cleaning (Indian dataframe)

Fig 3.3 shows the dataframe used for showing the data of all the parts of the world, using English as the medium of expressing emotions. Whereas Fig 3.4 shows the dataframe when it is filtered for India specific tweets.

3.2.3 Design of Experiments

The analysis will be performed on the both the datasets which will involve calculating various measures. This will be done using inbuilt libraries and functions provided by Numpy and Pandas to be performed on the data frames.

3.2.3.1 Average Likes over the period

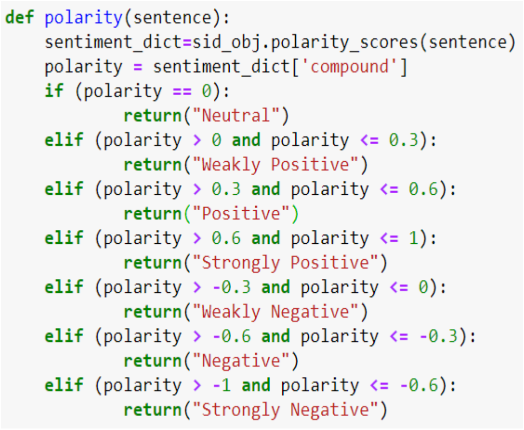
Likes show similarity in thoughts. People reading tweets online if relates to the tweet of the other person, they will definitely drop a like. This is one of the amazing feature of social media platforms and a must have too. Since I have data for each date, I will sum up the likes for each day and calculate the average/mean by using appropriate functions. This will create a time series plot of the above calculated.

3.2.3.2 Average Re-tweets over the period

Re-tweets are sharing a tweet and expressing your thoughts on that tweet. This helps one to know what people think and react to views of other people. There can be re-tweets of the already re-tweeted tweets; this is also counted but only as one single value. So similarly like I calculated the mean of likes, the same way mean retweets is calculated giving the time series plot which will be shown through one of the chart types.

3.2.3.3 Intensity Analysis

Every text has an intensity value, which signifies how good or bad the text is. I have used Vader sentiment analyser to calculate the intensity.



```
def polarity(sentence):
    sentiment_dict=sid_obj.polarity_scores(sentence)
    polarity = sentiment_dict['compound']
    if (polarity == 0):
            return("Neutral")
    elif (polarity > 0 and polarity <= 0.3):
            return("Weakly Positive")
    elif (polarity > 0.3 and polarity <= 0.6):
            return("Positive")
    elif (polarity > 0.6 and polarity <= 1):
            return("Strongly Positive")
    elif (polarity > -0.3 and polarity <= 0):
            return("Weakly Negative")
    elif (polarity > -0.6 and polarity <= -0.3):
            return("Negative")
    elif (polarity > -1 and polarity <= -0.6):
            return("Strongly Negative")
```

Fig 3.5 Method used to assign sentiment value to the text

It provides a polarity_score () method which takes in input the text of the tweet and the resultant is the polarity of the entire sentence, which is calculated word by word. I have divided the polarity into 7 different, with one of them being neutral and the other ranges between different types of positive and neutral values.

3.2.3.4 Polarity and Subjectivity

These two are one of the most important features of sentiment analysis. Polarity as discussed earlier is the intensity or we can call it the strength of the emotion one is showing through text which shows the behaviour of the person. Subjectivity on the other hand is personal opinion or view of a person regarding a particular issue or thing. It’s not necessary that a subjective sentence defines polarity or depicts a behaviour i.e. it can just be a normal statement made as a opinion.

I will be using Text Blob for this purpose; this takes input as the sentence for which the result is to be calculated. The function returns two things the first one is the polarity of the text and the second is the subjectivity. A mean of these values will be taken for each date and relevant plots will be made.

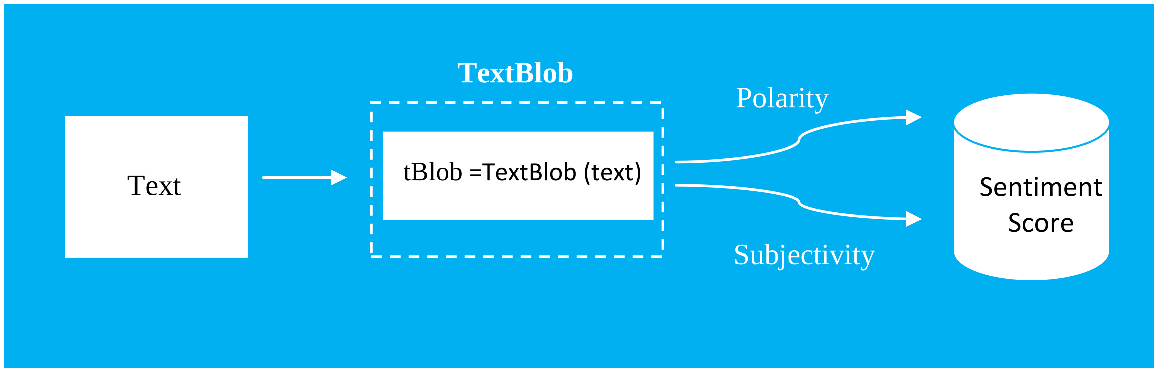


Fig 3.6 shows the internal working of the TextBlob method

Fig x.x shows the working of TextBlob method and what result its produces. The polarity score will have value ranging from [-1 to +1] and subjectivity score of the text ranges from [0 to +1].

3.2.3.5 Wordcloud

Each sentence has number of words and different words show different intensity and behaviour. As we have already calculated the type of polarity earlier, I will now scale those 7 types into just 3 types; i.e. positive, negative and neutral. Each of them will have a wordcloud showing different words that come in that category. Wordcloud shows all the words and the frequency is shown as the size of the respective word. Bigger the size of the word more frequent the word was used in the text. Also there could be stop words that could be repeated more than any other words, but these words don’t make any sense and doesn’t show the emotion. Hence, these words are removed initially before making the word cloud.

3.2.3.6 Implementing BERT

BERT stands for bi-directional encoding representation for transformer and this model is used for text classification, question answering and various other uses. It has contextual word embedding i.e. a word in a sentence may make it mean differently, and this depends on the words surrounding that word. It feeds all input at once therefore it can handle dependencies between words. It has two types bert-base and bert-large, the difference is that one uses 12 transformer encoders and the other uses 24 transformer encoders. Bert can be easily fine tuned to get desired results.

The steps that followed the implementation are creating a mask and the encoder representation of it using hugging face with pytorch. The data is divided into test and training set. Further the data is prepared for training using train tests split and this data is converted to torch tensors, which is the required data type for the model in process.

The next step is defining the batch size and creating tensors and iterators which are used in fine tuning of the Bert model and then followed by initialising the bert model. The last step is the Bert training loop where the model parameter is used to set the model into training mode and then into evaluation mode to get the validation accuracy.

3.3 System Requirement

With advancement in technologies high end processors are needed to run applications. The type of system required depends on the application you want to run. The following are the requirements necessary to develop a machine learning model.

3.3.1 Software Requirements

Python provides a vast range of libraries and frameworks which makes it easy for the developers to write concise and readable code. We have used Python 3.7 for coding and libraries such as Numpy, Pandas, and Scikit Learn.

Anaconda, which is open source software, used for developing ML and AI projects. It various packages useful for making an application, to further deploying it for production. Anaconda 2019.10 version was used which comes with Python 3.7 windows installer. You can choose from the two given options: Graphical Installer & Command Line Installer

Jupyter Notebook, which is one of the packages that comes with anaconda and is an Integrated Development Environment (IDE), made for programming in Python. Jupyter Notebook is a web application therefore, needs a web browser to run. It could be Google Chrome or Mozilla Firefox. It can be used for data cleaning and transformation as well as visualizing data using various inbuilt functions.

3.3.2 Hardware Requirements

Hardware as mentioned is an important requirement to develop anything. Computers nowadays come with hardware which is enough to build applications. Features like a running operating system with any processor should be available. Since applications can demand multiprogramming, the processor is advised to have 2 or more cores. A good memory space and a minimum RAM of 4 GB should be present to run software mentioned above.

3.4 Conclusion

After obtaining the desired datasets and filtering the content, I performed all the above mentioned operations on both the data frames. The result was values that were obtained from the analysis. These values will be helpful in plotting graphs and analyzing them accordingly.

CHAPTER 4

PERFORMANCE ANALYSIS

CHAPTER 4: Performance Analysis

4.1 Introduction

In this section I will be use the values from the obtained from the operations performed on the dataset, to plot various types of graphs and conclude what does the graph signifies. Since graphical insights are the best to conclude things and also one could easily make out what's going on by just looking at them.

4.2 Likes and Retweets over the period

As discussed earlier how likes matter a lot in the current scenario of social media platforms. I calculated the mean likes for each day, which resulted in a period of 80 days. This was done on both the dataframes. Fig 4.1 shows the mean likes over the period for the entire world and Fig 4.2 shows the mean likes for just India. Since the outbreak of corona virus in Wuhan people around the world got curios about what this virus is, how is it affecting people at such a large scale. We can see from the plot that for the world dataset two dates 29Jan and 9Mar have seen large hikes in the like and retweets.

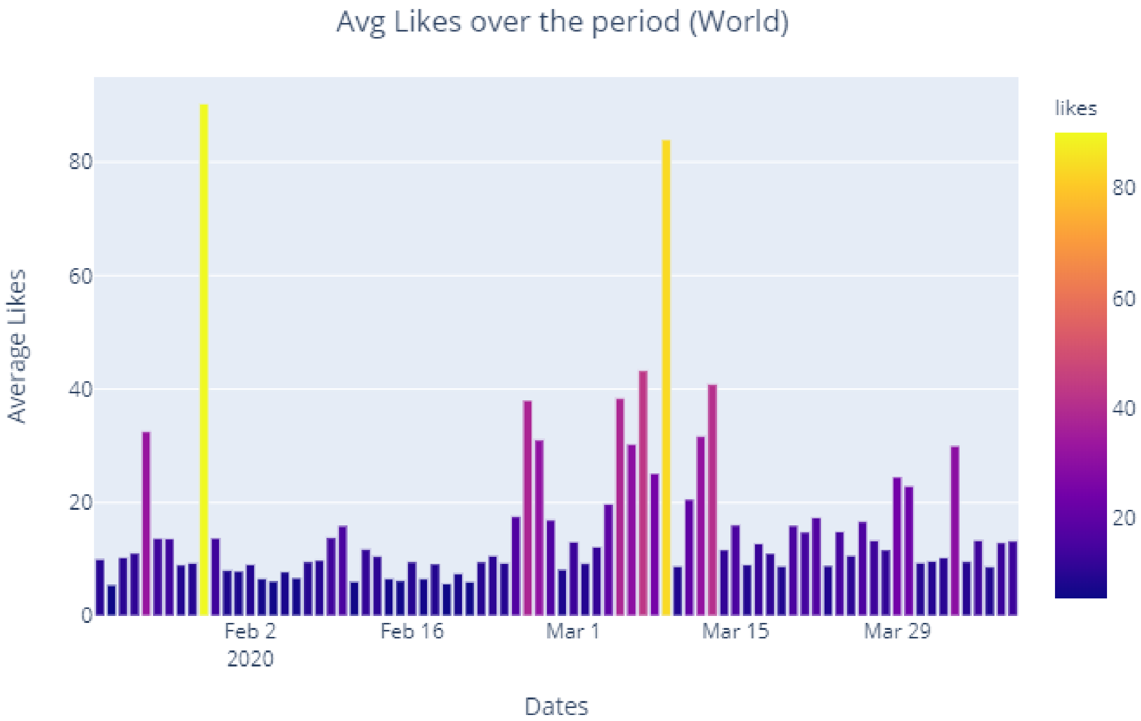


Fig 4.1 shows the plot for mean likes over the period for the world tweets dataframe

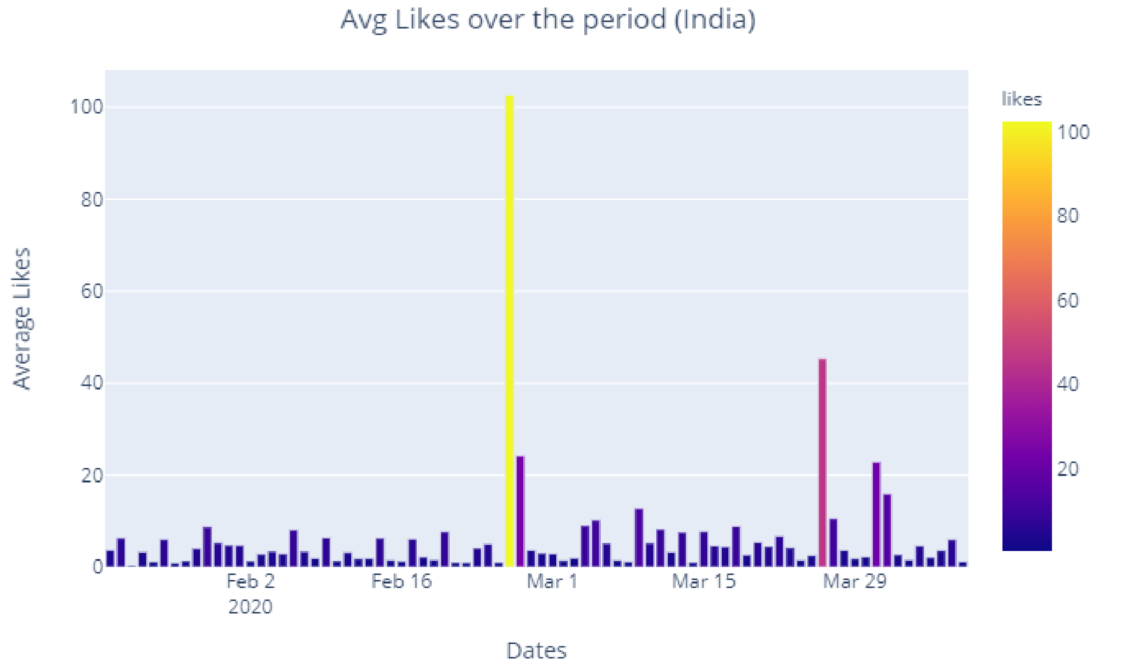


Fig 4.2 shows the plot for mean likes over the period for the Indian tweets dataframe

This is because this was the stage when cases in china were rising and deaths cases were around 200.Students studying in China and people from other countries were to be evacuated from there. US also sent flights to bring back citizens from Wuhan. Also there were delays seen in the process. Therefore, people were more active about the situation and more interaction followed.

Coming to India the plot shows 26Feb and 26 Mar as the two significant dates for the rise

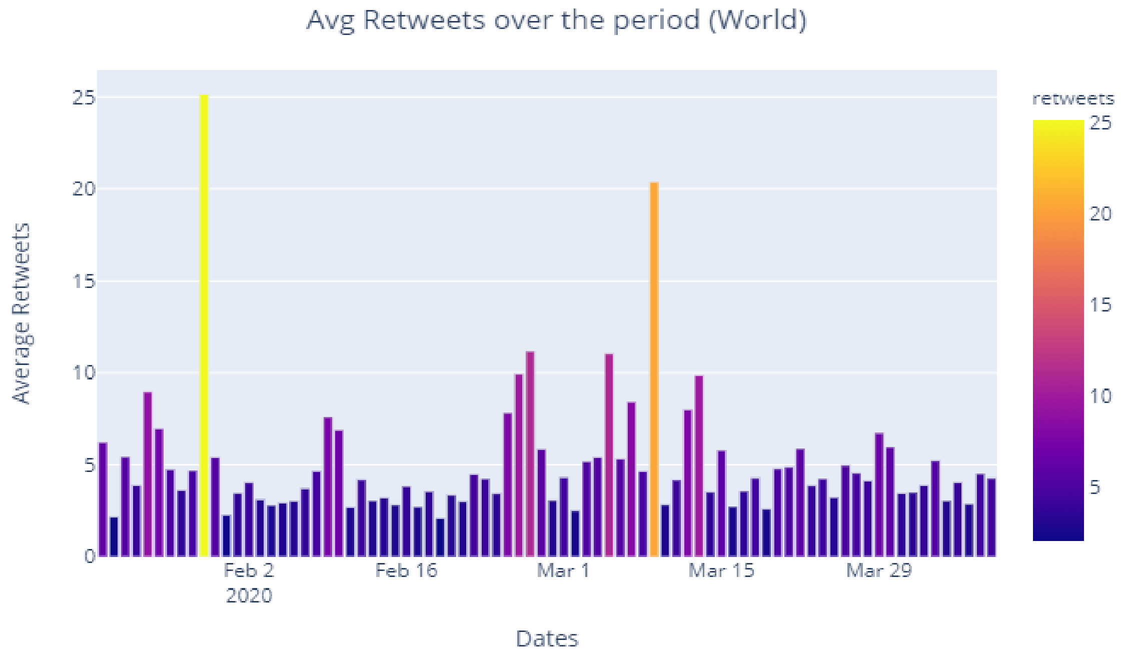


Fig 4.3 shows the plot for mean tweets over the period for the world tweets dataframe

in the number of likes and retweets. This is the time when India after clearance from the government decided to send military aircraft to deliver relief goods and bring back citizens from Wuhan. The next tower in March shows the time when a three week lockdown was announced over the entire country; this will definitely have reactions hence you can seek the hike.

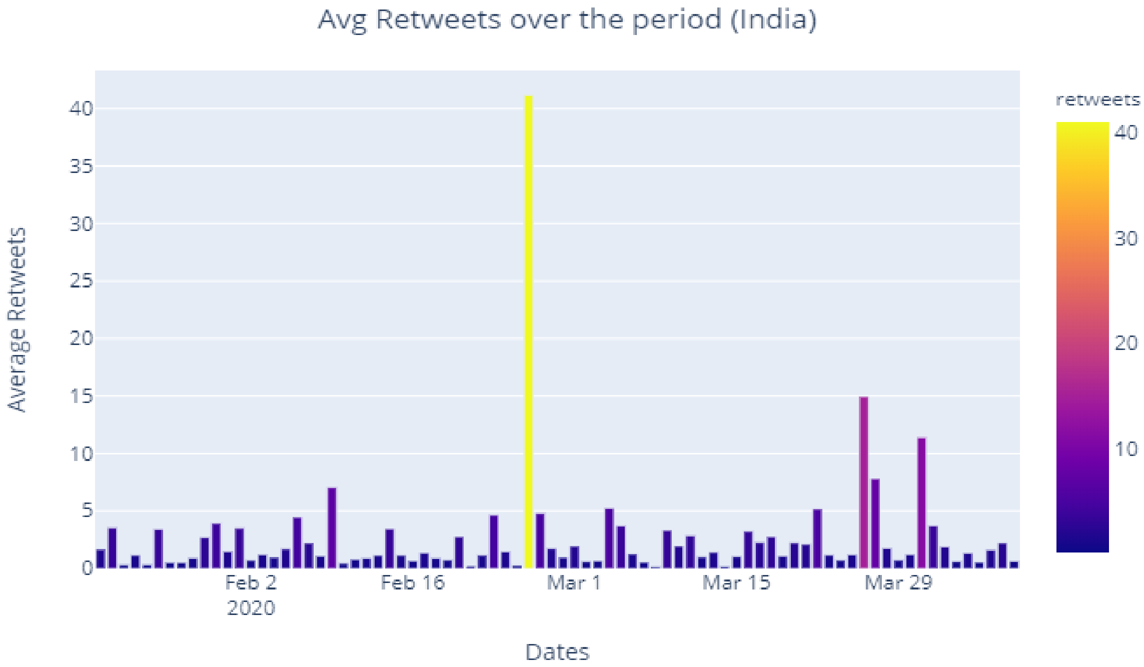
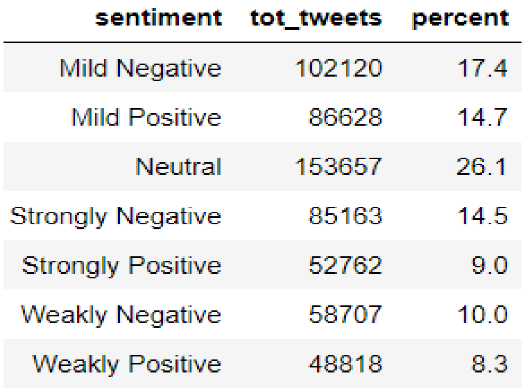


Fig 4.4 shows the plot for mean retweets over the period for the Indian tweets dataframe

4.3 Intensity Analyzer

I discussed about what process I did to derive intensity values into 7 different types. These values ranged between -1 to +1 and beyond that are used as strong intensity values.

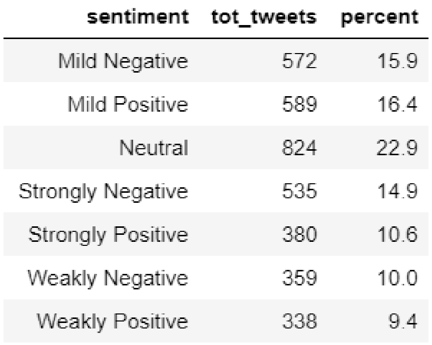
This shows the polarity as in how will be the statement judged if a third person is reading it. The seven groups were namely neutral, weakly positive, mild positive, strongly positive, weakly negative, mild negative and strongly negative. These were determined earlier and I decided to make a pie chart for the resultant percentage values.



sentiment	tot_tweets	percent
Mild Negative	102120	17.4
Mild Positive	86628	14.7
Neutral	153657	26.1
Strongly Negative	85163	14.5
Strongly Positive	52762	9.0
Weakly Negative	58707	10.0
Weakly Positive	48818	8.3

Fig 4.5 shows the plot percent values of the sentiment group for World dataframe

Fig 4.5 shows the values obtained from the world data frame in numbers and percentage for all the seven groups and values for single 3 groups were also calculated and Fig 4.6 shows the values for the Indian dataset. I used SunBurst pie chart to show the results in a more understandable manner.



sentiment	tot_tweets	percent
Mild Negative	572	15.9
Mild Positive	589	16.4
Neutral	824	22.9
Strongly Negative	535	14.9
Strongly Positive	380	10.6
Weakly Negative	359	10.0
Weakly Positive	338	9.4

Fig 4.6 shows the plot percent values of the sentiment group for Indian tweets dataframe

The plots made showed segments of each of the main type. Fig 4.7 shows the pie plot of the world dataset and Fig 4.8 shows the output percentage for the Indian dataset respectively.

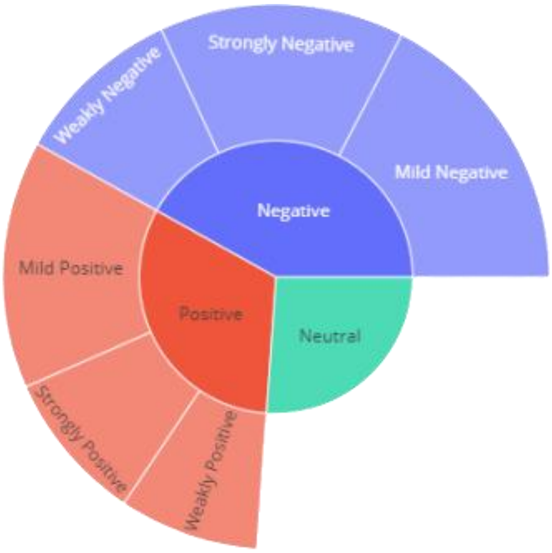


Fig 4.7 shows the pie plot for the sentiment group for world dataset

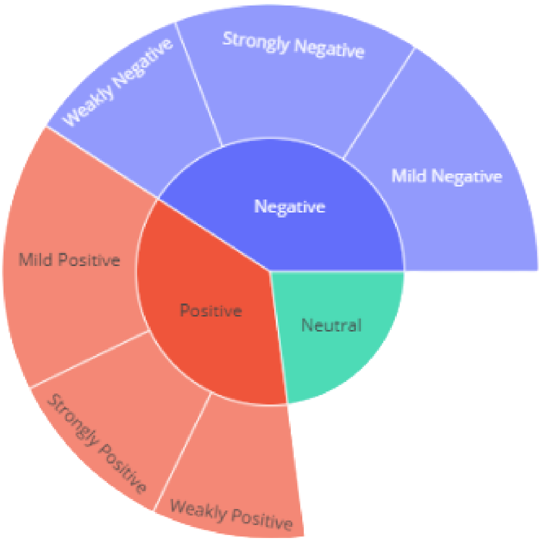


Fig 4.8 shows the pie plot for the sentiment group for Indian dataset

Firstly if we talk about the neutral tweets, the world dataset had more neutral sentiments compared to India. But the strongly negative tweets were more in the Indian dataframe; this was due to the hate for the corona virus in the initial stages. Along with this it also saw more number of strongly positive tweets as compared to the world data frame. This was mainly because of the actions taken by the government for helping out people and handling the situation.

4.4 Polarity and Subjectivity

This experiment was to check the sentiments and its intensity all over the period of 80 days. The mean was taken to make 80 unique records for each date. Fig 4.9 shows the Polarity over the period for the world dataset and Fig 4.10 shows the mean values for the Indian dataset. It can be seen that the mean value for the world sentiments did not drop below zero and was mostly in range 0.03 to 0.07. Whereas for the Indian dataset the values dropped below zero at various intervals. During the lockdown phase and while the virus was spreading the values saw the most negative polarity.

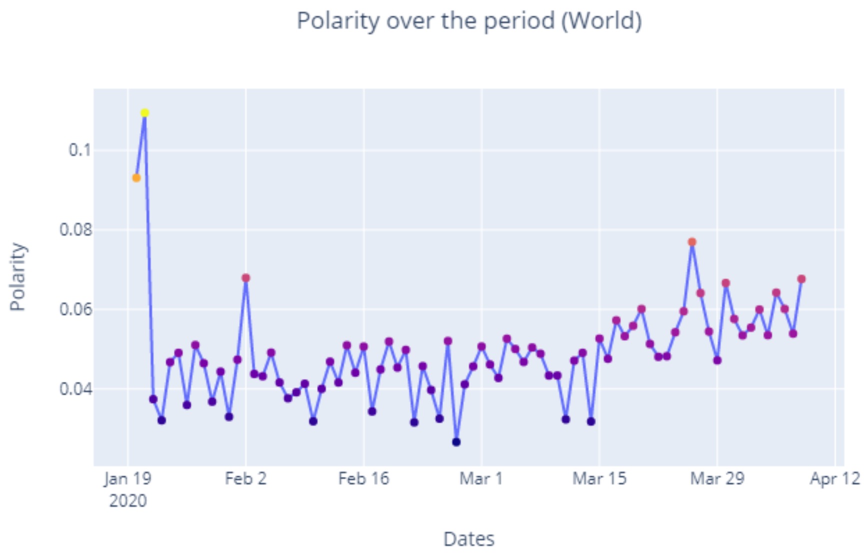


Fig 4.9 shows the scatter plot for mean polarity over the period for the world dataset

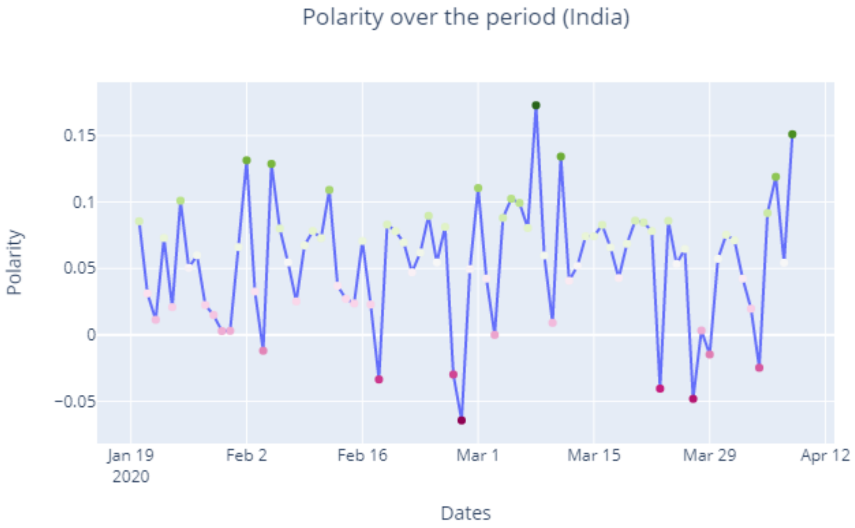


Fig 4.10 shows the scatter plot for mean polarity over the period for the Indian dataset

After plotting polarity, I used the mean values of subjectivity and plotted it using the same scatter plot. These types of plots change color of the markers according to the values, y axis in these four cases. Fig 4.11 shows the mean subjectivity values of the World dataset and Fig 4.12 shows the values for the Indian dataset. We can see that the result in the first case is compact within a small limit and makes two segments gradually increasing in March and then fluctuates normally. But in the second case it varies in much larger limit and it can be seen how instantly it changes the next day in some of the cases.

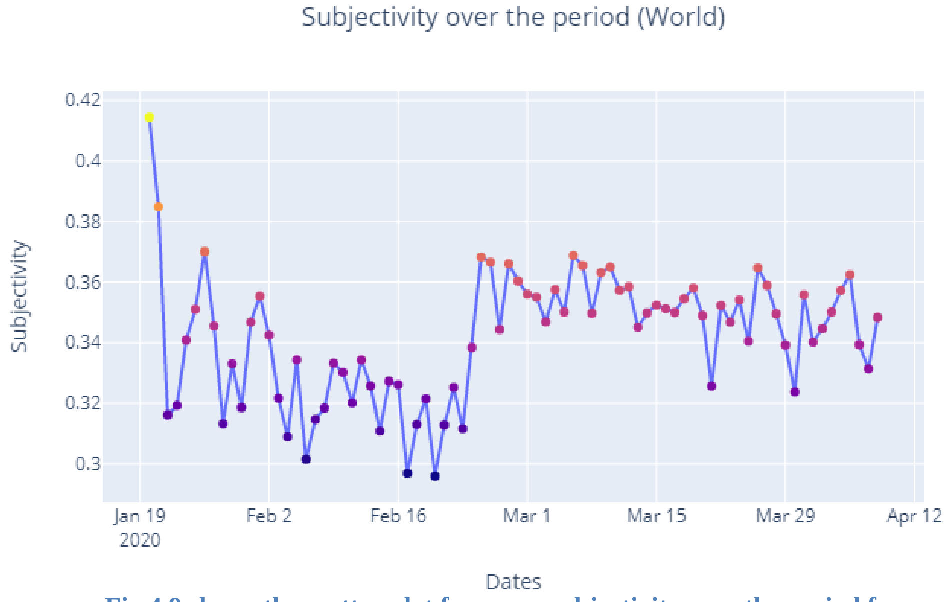


Fig 4.9 shows the scatter plot for mean subjectivity over the period for the world dataset

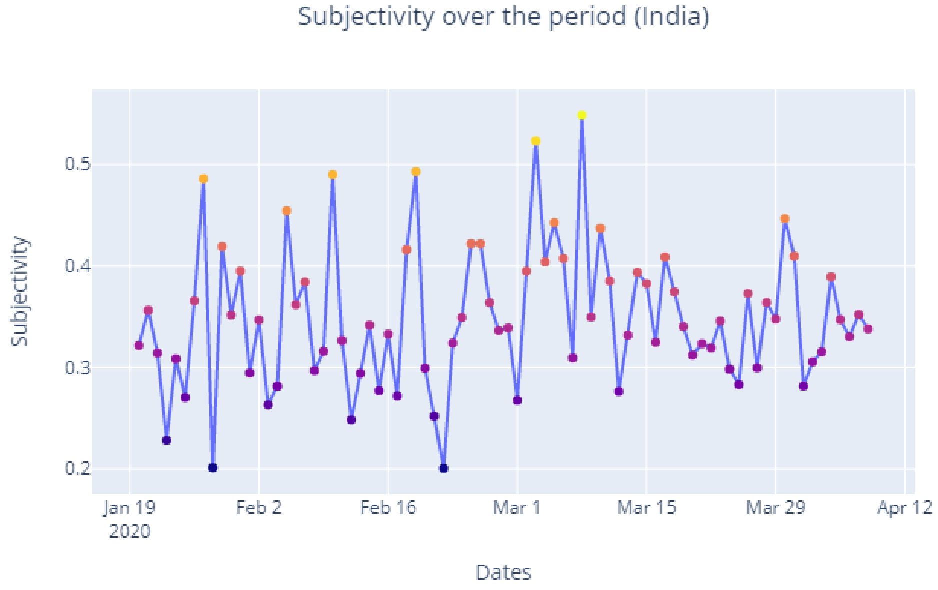


Fig 4.12 shows the scatter plot for mean subjectivity over the period for the Indian dataset

4.5 WordCloud Representation

After plotting all the numeric information gathered. I used the text to plot wordclouds for all the three groups namely neutral, positive and negative. Three different data frames were created using the sentiment value calculated previously. Tweet text was used to plot these clouds and more stop words were added to filter the words.

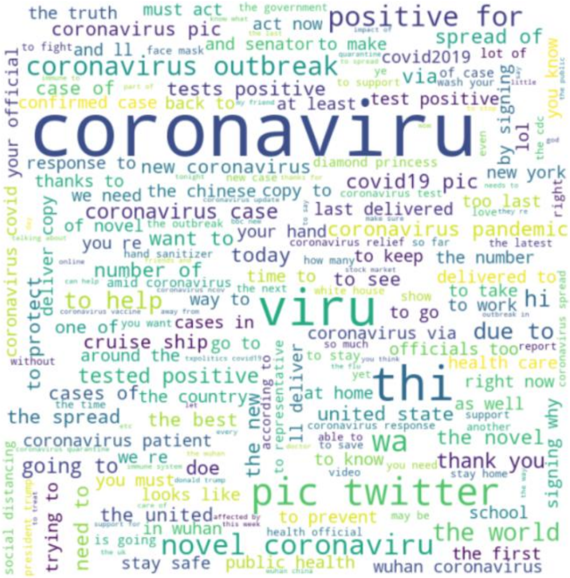


Fig 4.13 Positive word cloud (world data)

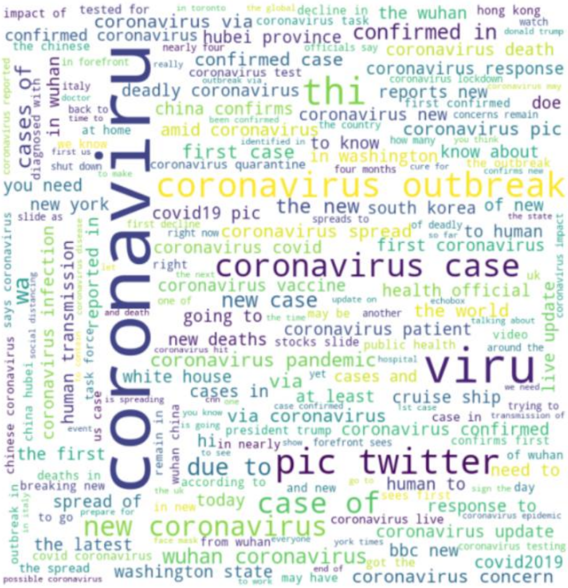


Fig 4.14 Neutral word cloud (world data)

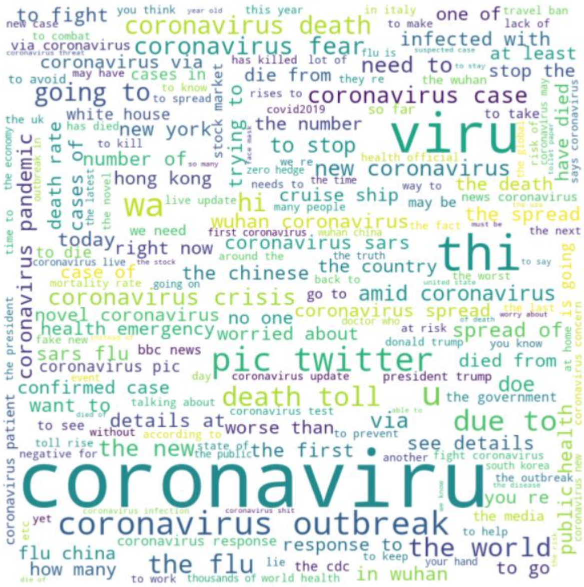


Fig 4.15 Negative word cloud (world data)

Fig 4.13, 4.14, 4.15 shows the world cloud made for the positive, neutral and negative text sentiments for the world data frame. We can see that terms like flu, corona virus and many other words which are neither of the three are seen in all the word clouds. In the positive word space words like health, commodities, crew, charge, vaccine, god, save, hospital etc are used which signifies what actions were taken during the outbreak. Later in

Neutral space words like new cases, Wuhan, treatment, Chinese, Germany, Italy, confirmed, via etc are which are commonly the opinions and reactions to the spreading in various parts of the world. Coming to the negative space words like pandemic, death, fear etc are used. This shows the issues people are facing during the times of the crisis. Economy is also seen in the cloud, since the crises shut down many small and major businesses all around the world, people and even the government faced economic issues. Words showing actions like, travel ban, fight back, infected with show a lot of emotions. People were very scared about the travel ban as some were restricted to stay at work places and couldn’t go home. Stock market on the other hand saw a decrease and imbalance in the output the generally produced. Also it is seen that people used the word china and corona virus a lot, maybe they have hate or angry about china being so careless about the situation; words like worse is used as the death rates started increasing in various parts of the world.

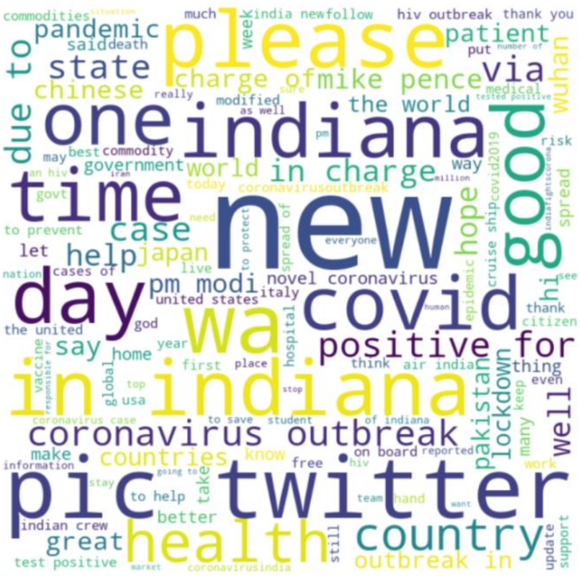


Fig 4.16 Positive word cloud (Indian data)

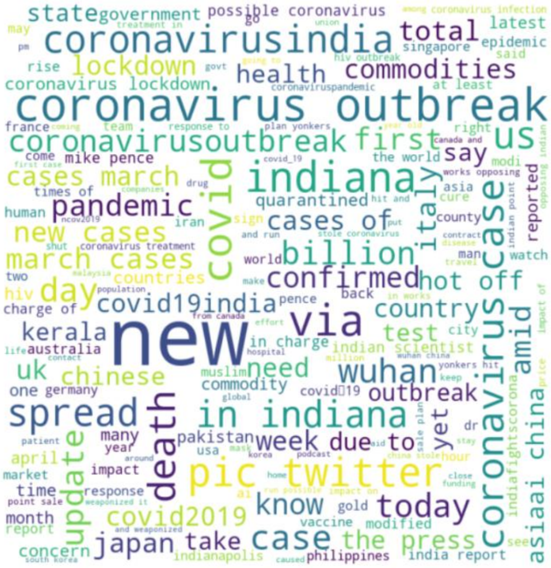


Fig 4.17 Positive word cloud (Indian data)

Fig 4.16, 4.17, 4.18 shows the world cloud made for the positive, neutral and negative text sentiments for the world data frame. India had a mixed reaction about the outbreak of the virus. Words repeated in the plots can be seen. Government has been constantly questioned about the situation. Every time the rate increases we can see tweets with words like new, infected, confirmed, deaths etc. Lockdown is one of the major terms used and had mixed reactions. Many people supported the decision but there were people who didn't get along with the idea. You will be already aware about the results after the lockdown was declared and how it affected the poor. Our PM had many question from all the people too. With increase in cases around the world people predicted how India will be at risk in the coming times and how this will spread and affect the people and the economy. Demand being used because people wanted their family members back home, who were stuck in foreign countries or studying abroad. Wuhan city was constantly targeted for the spreads of the virus. Many started working in the process of vaccination

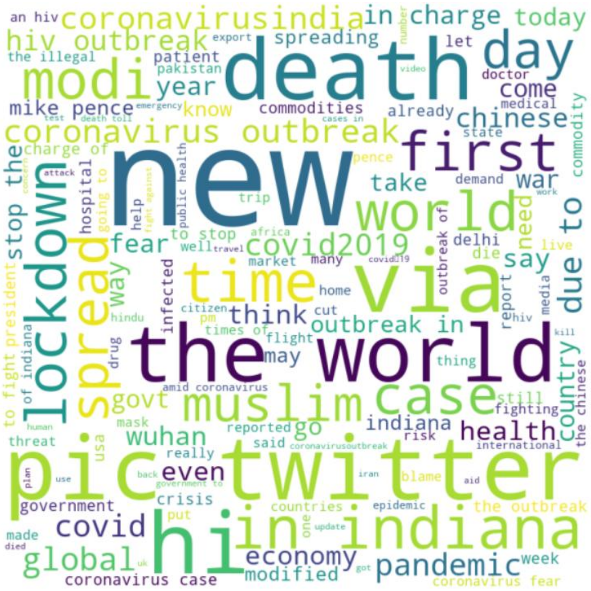
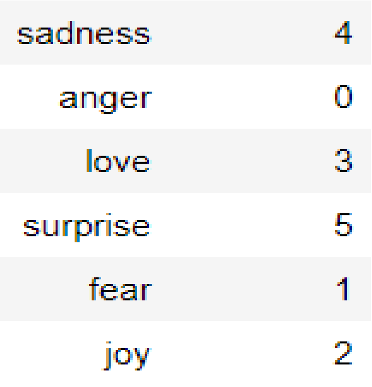


Fig 4.18 Positive word cloud (Indian data)

And building a medicine for the virus, HIV is therefore used because there were findings that medicine used to treat HIV patients can be used to moderate the affect of corona too. But since now no certain vaccine is made to cure the virus.

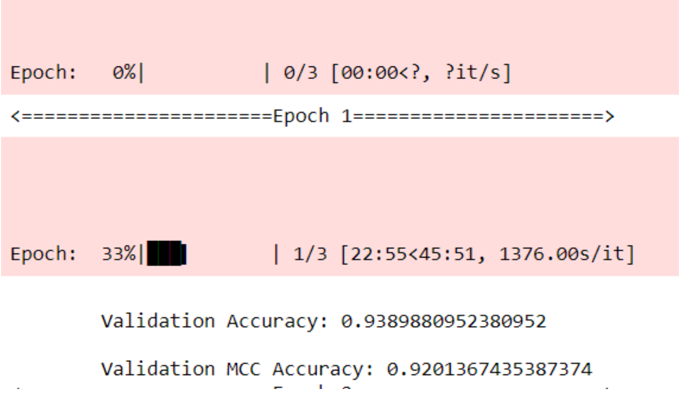
4.6 BERT Implementation

As discussed how Bert is fine tuned and how data is prepared for training. I implemented as per an article [11]. I tool pre labelled data from a Github repository who worked with the similar data. Fig 4.19 shows the 6 emotion groups and the labels encodes used for



sadness	4
anger	0
love	3
surprise	5
fear	1
joy	2

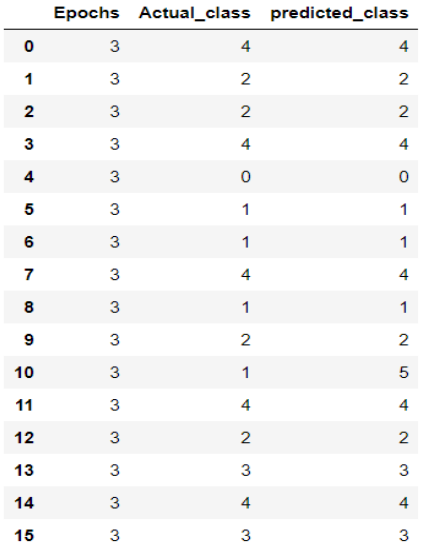
Fig 4.19 Emotions and labels used for Bert classification



```
Epoch:   0%|          | 0/3 [00:00<?, ?it/s]
<======================Epoch 1======================>

Epoch:  33%|███       | 1/3 [22:55<45:51, 1376.00s/it]

        Validation Accuracy: 0.9389880952380952

        Validation MCC Accuracy: 0.9201367435387374
```

Fig 4.20 Emotions and labels used for Bert classification

them using the label encoder. Fig 4.20 shows the output from the training loop implemented with the validation accuracy as 0.9389. Lastly Gig 4.21 shows the label encodes for actual and predicted values. We can see that for serial number 10 the actual label is 1 but the predicted label is 5. Rest all are predicted correctly and hence the resulted accuracy. One can also show the training operations like the training loss through line chart by just printing it out.



	Epochs	Actual_class	predicted_class
0	3	4	4
1	3	2	2
2	3	2	2
3	3	4	4
4	3	0	0
5	3	1	1
6	3	1	1
7	3	4	4
8	3	1	1
9	3	2	2
10	3	1	5
11	3	4	4
12	3	2	2
13	3	3	3
14	3	4	4
15	3	3	3

Fig 4.21 metrics for the implementation of the Bert model

4.7 Conclusion

After performing all the experiments mentioned this project work comes to and end. The operations were executed successfully and I got to learn about Bert which is a new hot topic in the field of sentiment analysis. Later I will discuss about the conclusions made and the future scope of the work carried out.

CHAPTER 5

CONCLUSION & FUTURE WORK

Chapter 5: Conclusion and future work

5.1 Conclusion

The worked carried out during the process started with introduction to sentiment analysis and the talking about the flow of the project report. Later I did the literature survey about the existing work on sentiment analysis and the techniques used to carry out the analysis. Further in the third module, I talked about preparing the dataset after scrapping and cleaning. The two obtained datasets after required filtering are the main components of the project. All the techniques and operations that were to be performed on these datasets were defined and the results were stored in dataframes. In the fourth module, the results obtained from the operations were plotted as graphical representations. Plots like mean likes and retweets, polarity pie chart, average polarity and subjectivity and world cloud for different sentiment groups were made. Then I applied Bert model for emotion classification by using the labelled in the training phase.

The conclusions made from all the plots is that throughout the period of the dataset, likes and tweets increased on particular dates at a very high level compared to other dates. Comparing the pie plots of both the datasets, Indians had more percentage of strongly negative as well as strongly positive tweets and on the other hand tweets from all around the world had more neutral tweets. But for all data combined both the scenarios had a higher percent of negative tweets showing that it affected the people’s emotions and thus they had negative thoughts about the situation. Talking about the polarity the tweets from around the world showed positive mean polarity throughout the period but India constant ups and down were seen. Mainly during the start of the lockdown until the first week it stayed negative because of the difficulties faced in the lockdown and the travel ban imposed suddenly. But by the end of the week it increased and at the end of the period it stayed positive. Subjectivity on the other hand did not cross 0.5 in case of world dataset but it did for the second case. Later talking about the word cloud for both the cases we can see different words being used frequently over the period. Finally implementing the BERT model;

5.2 Future Work

The above work made use of the twitter scrapper which does filter location by itself. Hence making use of a better API will help get more tweets. Carrying out sentiment analysis for such a pandemic is very important as it plays with mind as well. Taking in consideration the state of mental health, it is the most important thing nowadays and shouldn’t be left as a disability. If real time analysis can be performed, even the government officials can take care of what’s happening in their country. There could be cases where people are talking about their depression or maybe hate comments, but no one really judges the emotion. Through these types of analysis this can be taken care of. Hence future work would be just taking care of all the above aspects in a more real time manner.

References

1. Mark Holmstorm, Dylan Liu, Christopher Vo, “Machine Learning Applied to Weather Forecasting”, Stanford University, December 15 , 2016

2. Pang, Bo and Lee, Lillian. 2008. Analysis mining opinion sentiment. Journal of Foundations and Trends in Information Retrieval, 2, 1–135.

3. Abdullah Alsaeedi1, Mohammad Zubair Khan, “A Study on Sentiment Analysis Techniques of Twitter Data”, IJACSA Vol. 10, No. 2, 2019.

4. Dave, S. L. K. and Pennock, D. M. 2003. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews, Proceedings of the 12th International Conference on World Wide Web. New York, NY, USA, 519-528.

5. Agarwal A., Xie B., Vovsha I., Rambow O., Passonneau R. 2011. Sentiment Analysis of Twitter Data. Proceedings of the Workshop on Languages in Social Media. Stroudsburg, PA, USA, 30-38.

6. Kim, S.-M. and Hovy, E. 2006. Automatic Identification of Pro and Con Reasons in Online Reviews. Proceedings of the COLING/ACL Main Conference Poster Sessions, Stroudsburg, PA, USA, 483–490.

7. Esteban Castillo, Ofelia Cervantes, Darnes Vilarino, David Baez and Alfredo Sanchez, UDLAP: Sentiment Analysis Using a Graph Based Representation 1Universidad de las Americas Puebla ´ Department of Computer Science, Electronics and Mechatronics, Mexico

8. Ahmed Imran KABIR, Ridoan KARIM, Shah NEWAZ3 , Muhammad Istiaque HOSSAIN, "The Power of Social Media Analytics: Text Analytics Based on Sentiment Analysis and Word Clouds on R", Informatica Economică vol. 22, no. 1/2018

9. Hamid Bagheri ; Md Johirul Islam, “Sentiment analysis of twitter data” Computer Science Department Iowa State University

10. Alexander Pak, Patrick Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", Universit´e de Paris-Sud

11. Michel Kana , BERT for dummies — Step by Step Tutorial

Project Report Undertaking

I Mr. Rahul Rana, Roll No. 161214 Branch CSE has done my internship with Infosys from 22 Feb 2020 to 22 March 2020 to.

As per procedure I have to submit my project report to the university related to my work that I have done during this internship.

I have compiled my project report. But due to COVID-19 situation my project mentor in the company is not able to sign my project report.

So I hereby declare that the project report is fully designed/developed by me and no part of the work is borrowed or purchased from any agency. And I'll produce a certificate/document of my internship completion with the company to TnP Cell whenever COVID-19 situation gets normal.

Signature______________________

Name: Rahul Rana

Date: 22 Jun 2020

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

PLAGIARISM VERIFICATION REPORT

Date: 14 JULY 2020

Type of Document (Tick): PhD Thesis | M.Tech Dissertation/ Report | B.Tech Project Report | Paper

Name: RAHUL RANA **Department:** CSE **Enrolment No** 161214

Contact No: 9882982867 **E-mail:** jr.rahul07@gmail.com

Name of the Supervisor: Dr Mrityunjay Singh
Title of the Thesis/Dissertation/Project Report/Paper (In Capital letters): SENTIMENT ANALYSIS OF COVID19

UNDERTAKING

I undertake that I am aware of the plagiarism related norms/ regulations, if I found guilty of any plagiarism and copyright violations in the above thesis/report even after award of degree, the University reserves the rights to withdraw/revoke my degree/report. Kindly allow me to avail Plagiarism verification report for the document mentioned above.

– Total No. of Pages = 44
– Total No. of Preliminary pages = 8
– Total No. of pages accommodate bibliography/references = 1

(Signature of Student)

FOR DEPARTMENT USE

We have checked the thesis/report as per norms and found **Similarity Index** at 3 (%). Therefore, we are forwarding the complete thesis/report for final plagiarism check. The plagiarism verification report may be handed over to the candidate.

(Signature of Guide/Supervisor) **Signature of HOD**

FOR LRC USE

The above document was scanned for plagiarism check. The outcome of the same is reported below:

Copy Received on	Excluded	Similarity Index (%)	Abstract & Chapters Details	
	• All Preliminary Pages • Bibliography/ Images/Quotes • 14 Words String		Word Counts	
			Character Counts	
Report Generated on		**Submission ID**	Page counts	
			File Size	

Checked by
Name & Signature **Librarian**

……

Please send your complete Thesis/Report in (PDF) & DOC (Word File) through your Supervisor/Guide at plagcheck.juit@gmail.com