Twitter Sentiment Analysis on Covid-19

A Project Report Submitted by

Sneh Singh (191B262)

Vanama Yaswanth (191B278)

Vidhi Mathur(191B282)

In partial fulfilment for the award of the degree

of

BACHELOR OF TECHONOLOGY IN

COMPUTER SCIENCE AND ENGINEERING

At

Jaypee University of Engineering and Technology,

Guna, Madhya Pradesh ,473226

DECLARATION

We hereby declare that the work reported in 6th semester Minor project entitled "Twitter Sentiment Analysis on Covid-19", in partial fulfilment for the award of the degree of B.Tech. submitted at Jaypee University of Engineering and Technology, Guna, as per the best of our knowledge and belief there is no infringement of intellectual property rights and copyright. In case of any violation, we will solely be responsible.

Signature of Student

Sneh Singh (191B262)

Vanama Yaswanth (191B278)

Vidhi Mathur(191B282)

Department of Computer Science and Engineering,

Jaypee University of Engineering and Technology,

Guna ,473226

Date:

CERTIFICATE

This is to certify that the project titled "**Twitter Sentiment Analysis on Covid-19**" is the bonafide work carried out by **Sneh Singh**, **Vanama Yaswanth** and **Vidhi Mathur**. We are students of B.Tech (CSE) at Jaypee University of Engineering and Technology Guna (MP) during the academic year 2021-2022 in partial fulfilment of the requirements for the award of the degree Of Bachelor of Technology (Computer Science and Engineering) and that the project has not formed the basis for the award previously of any other degree, diploma, fellowship or any other similar title.

Signature of Guide

Department of Computer Science and Engineering,
Jaypee University of Engineering and Technology,
Guna, 473226
Date:

ACKNOWLEDGEMENT

We take this opportunity to express our deep gratitude and most sincere thanks to our project supervisor Dr. Ajay Kumar and project coordinator Dr. Dinesh Kumar Verma for giving the most valuable suggestions, helpful guidance and encouragement in the execution of this project work. We would like to thank our supervisor and coordinator for guiding us. Last but not the least we are grateful to all the team members of "Twitter Sentiment Analysis of Covid-19".

Sneh Singh (191B262)

Vanama Yaswanth (191B278)

Vidhi Mathur (191B282)

Department of Computer Science and Engineering,
Jaypee University of Engineering and Technology,
Guna, 473226

Date:

ABSTRACT

This project addresses the problem of sentiment analysis in twitter; that is classifying tweets according to the sentiment expressed in them: positive, negative or neutral. Twitter is an online micro-blogging and social-networking platform which allows users to write short status updates of maximum length 140 characters. It is a rapidly expanding service with over 200 million registered users, out of which 100 million are active users and half of them log on twitter on a daily basis - generating nearly 250 million tweets per day . Due to this large amount of usage we hope to achieve a reflection of public sentiment by analysing the sentiments expressed in the tweets. Analysing the public sentiment is important for many applications such as firms trying to find out the response of their products in the market, predicting political elections and predicting socioeconomic phenomena like stock exchange. The aim of this project is to develop a functional classifier for accurate and automatic sentiment classification of an unknown tweet stream.

We have chosen to work with twitter since we feel it is a better approximation of public sentiment as opposed to conventional internet articles and web blogs. The reason is that the amount of relevant data is much larger for twitter, as compared to traditional blogging sites. Moreover, the response on twitter is more prompt and also more general (since the number of users who tweet is substantially more than those who write web blogs on a daily basis).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

This project of analysing sentiments of tweets comes under the domain of “Pattern Classification” and “Data Mining”. Both of these terms are very closely related and intertwined, and they can be formally defined as the process of discovering “useful” patterns in large set of data, either automatically (unsupervised) or semiautomatically (supervised). The project would heavily rely on techniques of “Natural Language Processing” in extracting significant patterns and features from the large data set of tweets and on “Machine Learning” techniques for accurately classifying individual unlabelled data samples (tweets) according to whichever pattern model best describes them.

The features that can be used for modelling patterns and classification can be divided into two main groups: formal language based and informal blogging based. Language based features are those that deal with formal linguistics and include prior sentiment polarity of individual words and phrases, and parts of speech tagging of the sentence. Prior sentiment polarity means that some words and phrases have a natural innate tendency for expressing particular and specific sentiments in general. For example, the word “excellent” has a strong positive connotation while the word “evil” possesses a strong negative connotation. So, whenever a word with positive connotation is used in a sentence, chances are that the entire sentence would be expressing a positive sentiment. Parts of Speech tagging, on the other hand, is a syntactical approach to the problem. It means to automatically identify which part of speech each individual word of a sentence belongs to: noun, pronoun, adverb, adjective, verb, interjection, etc. Patterns can be extracted from analysing the frequency distribution of these parts of speech (ether individually or collectively with some other part of speech) in a particular class of labelled tweets. Twitter based features are more informal and relate with how people express themselves on online social platforms and compress their sentiments in the limited space of 140 characters offered by twitter. They include twitter hashtags, retweets, word capitalization, question marks, presence of URL in tweets, exclamation marks, internet emoticons and internet shorthand/slangs.

Classification techniques can also be divided into two categories: Supervised vs. Unsupervised and non-adaptive vs. adaptive/reinforcement techniques.

Supervised approach is when we have pre-labelled data samples available and we use them to train our classifier. Training the classifier means to use the pre-labelled to extract features that best model the patterns and differences between each of the individual classes, and then classifying an unlabelled data sample according to whichever pattern best describes it.

Unsupervised classification is when we do not have any labelled data for training.

There are several metrics proposed for computing and comparing the results of our experiments. Some of the most popular metrics include: Precision, Recall, Accuracy, F1-measure, True rate and False alarm rate (each of these metrics is calculated individually for each class and then averaged for the overall classifier performance.) A typical confusion table for our problem is given below along with illustration of how to compute our required metric.

There are several metrics proposed for computing and comparing the results of our experiments. Some of the most popular metrics include: Precision, Recall, Accuracy, F1-measure, True rate and False alarm rate (each of these metrics is calculated individually for each class and then averaged for the overall classifier performance.) A typical confusion table for our problem is given below along with illustration of how to compute our required metric.

	Machine says Yes	Machine says No
Human says Yes	Tp	fn
Human says No	Fp	tn

Table 1- Confusion matrix content

Tp=true positive Fn=false negative

Tn=true negative Fp=false positive

$$\textbf{Precision(P)} = \frac{tp}{tp+fp} \qquad \textbf{Recall(R)} = \frac{tp}{tp+fn} \qquad \textbf{Accuracy(A)} = \frac{tp+tn}{tp+tn+f+fp+fn}$$

2. OBJECTIVE

The main objective of this project is to perform sentiment analysis on Covid-19, what perception does the world hold about it, based on the tweets which are extracted from Twitter. Further, we extended the scope of this project to check the accuracy of Text blob, which is a Python library that returns the polarity and subjectivity of a sentence.

3. METHODOLOGY

To achieve this objective discussed above, following methodology is used:

- A thorough study of existing approaches and techniques in field of sentiment analysis.
- Collection of related data from Twitter with the help of Twitter API.
- Pre-processing of data collected from Twitter so that it can be fit for mining.
- To build a classifier based on different supervised machine learning techniques.
- Training and testing of built classifier using dataset collected from Twitter.
- Checking the accuracy of Textblob and comparing results of each classifier and visualising the results.

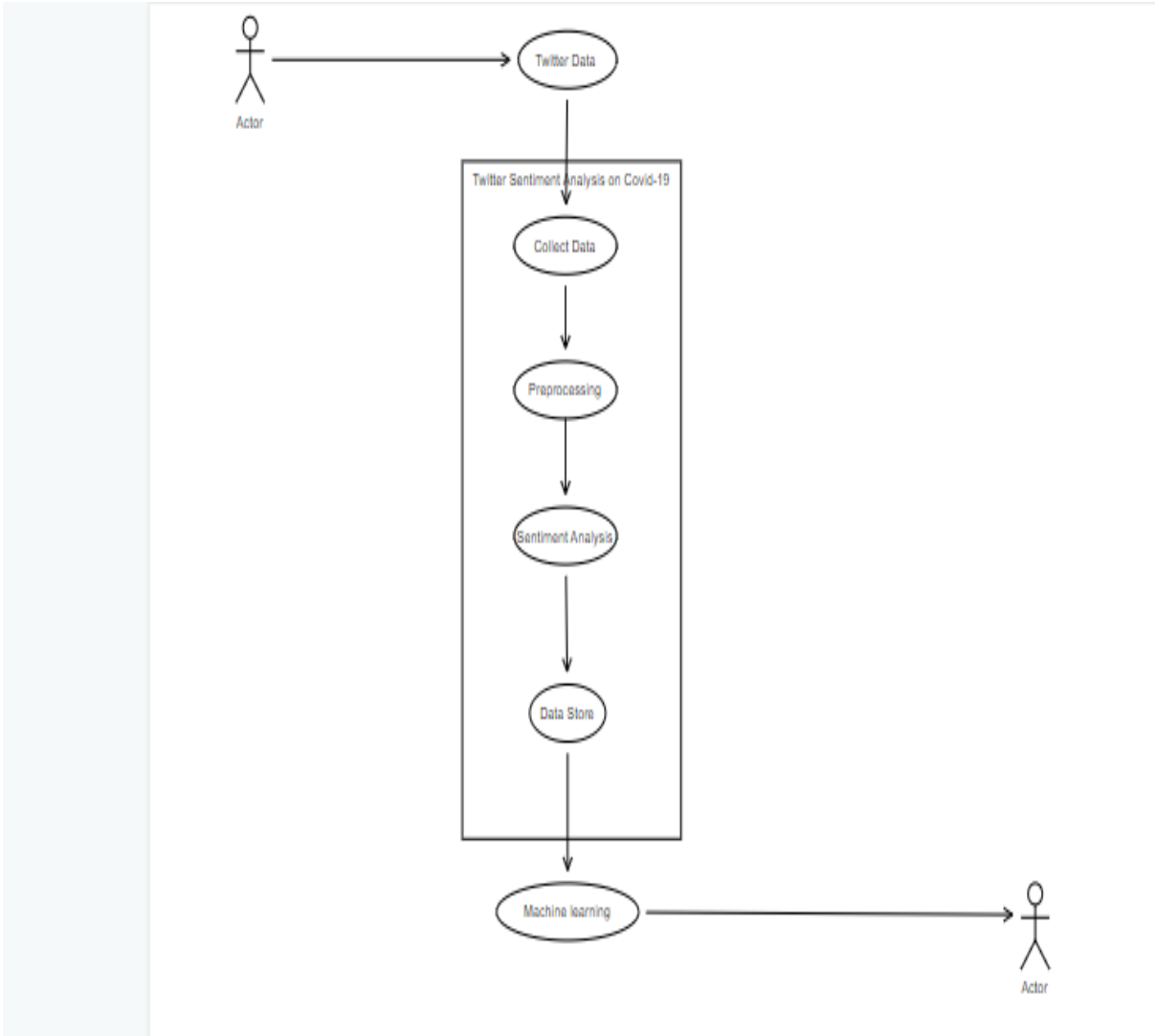


Figure 1: Use Case Diagram for the model

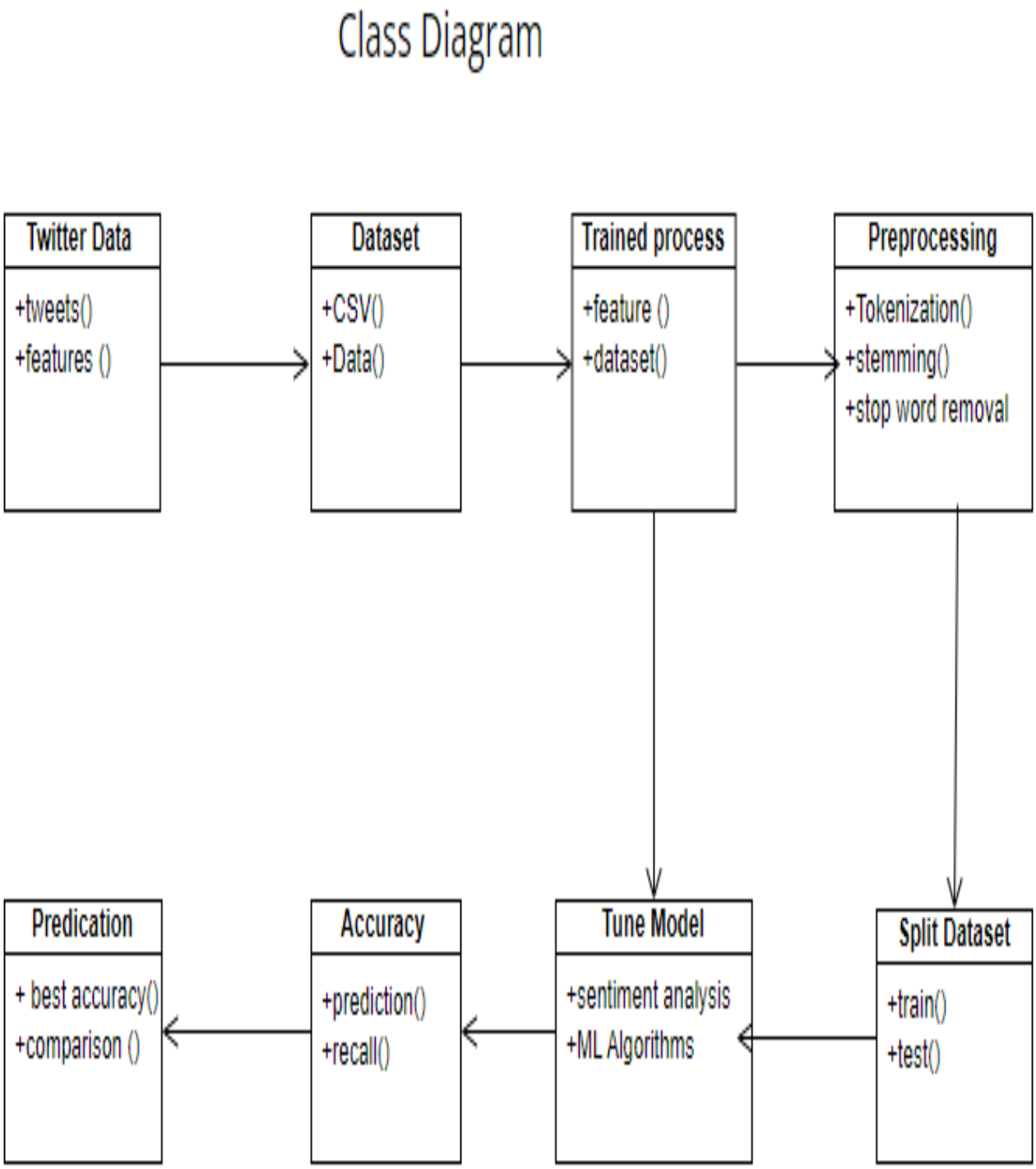


Figure 2: Class Diagram

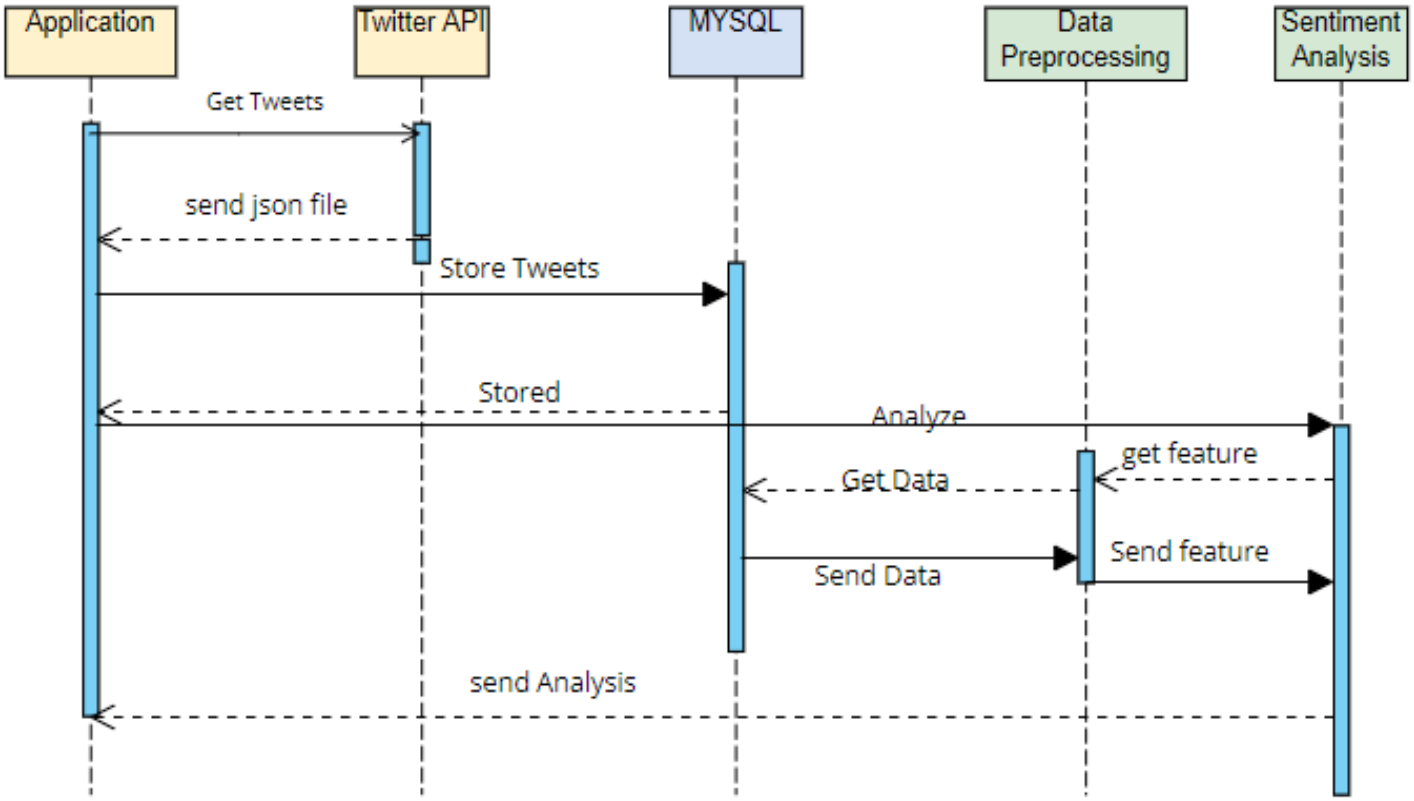


Figure 3: Sequence Diagram

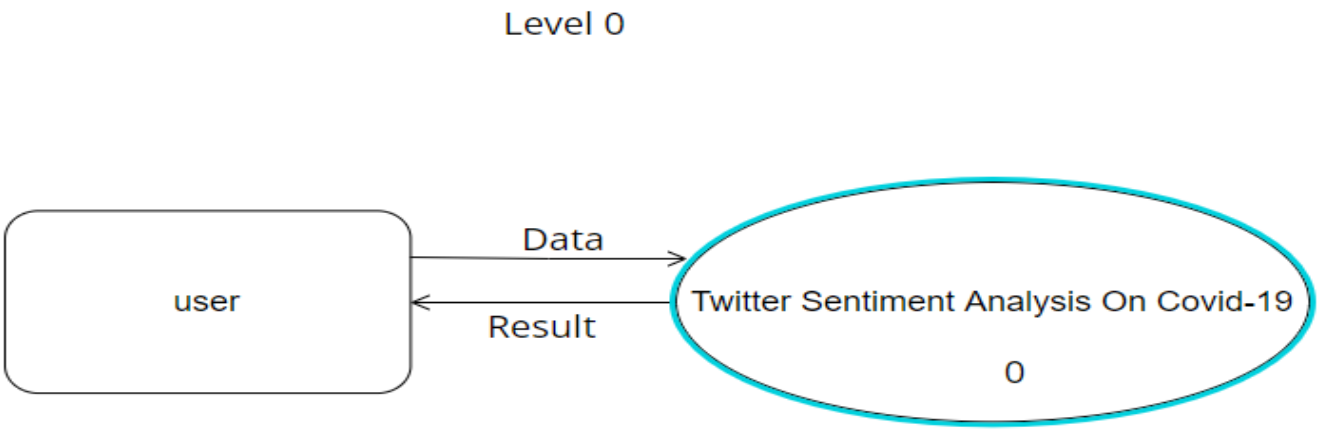


Figure 4.1: Data Flow Diagram Level 0

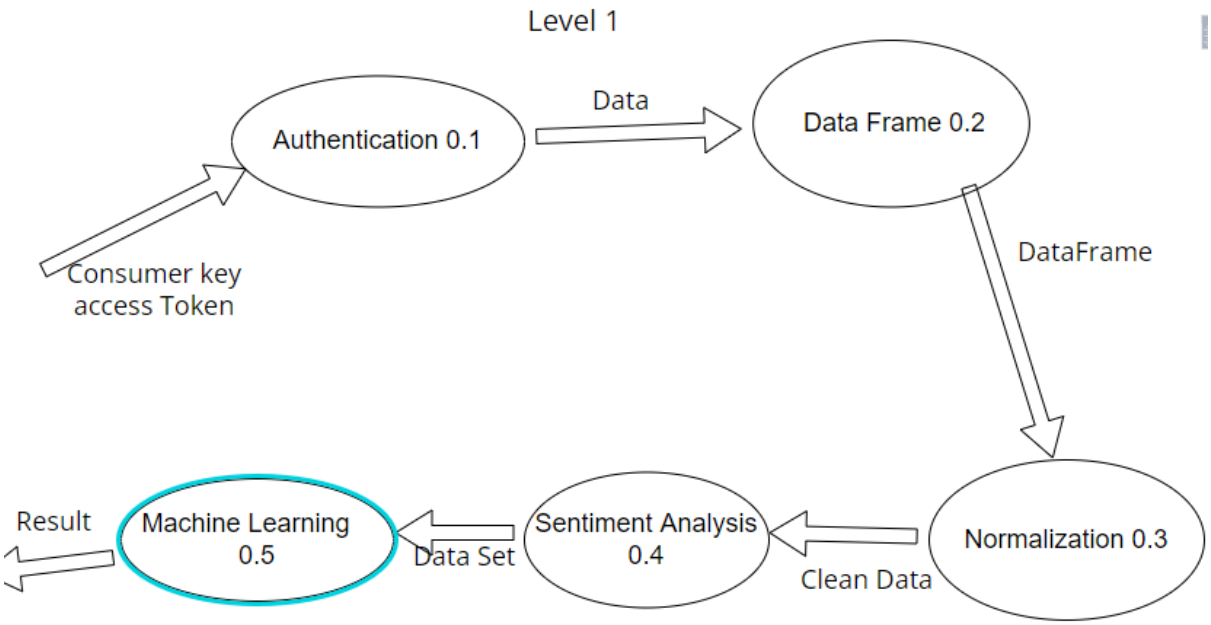


Figure 4.2: Data Flow Diagram Level 1

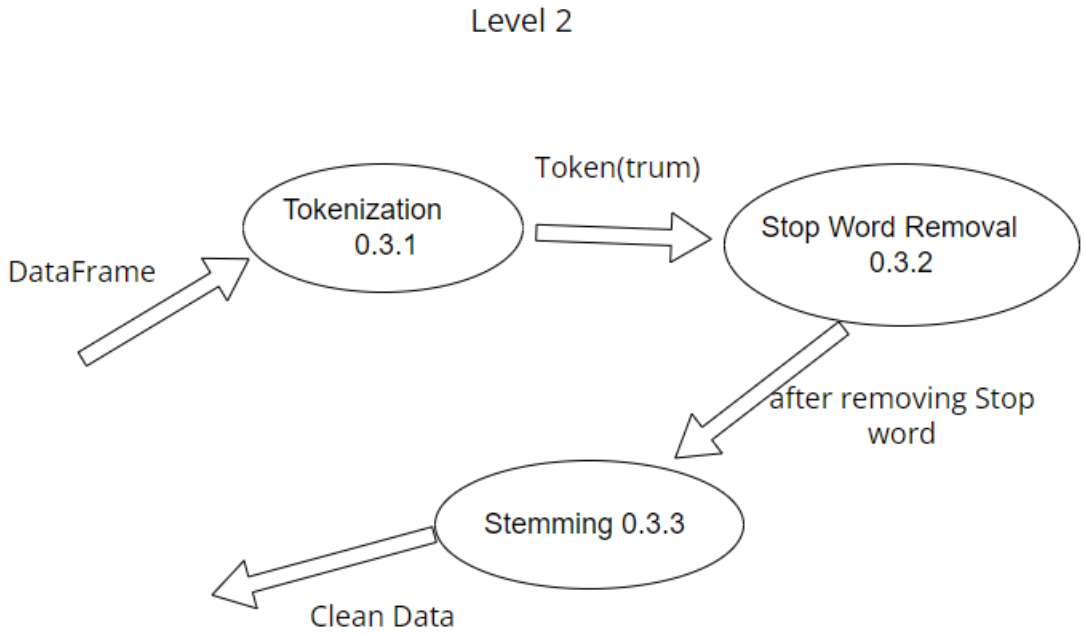


Figure 4.3 : Data Flow Diagram Level 2

3.1 Collection of Data

3.1.1 Twitter Data

Twitter allows users to collect tweets with the help of Twitter API. Twitter provides two kinds of APIs: REST API and Streaming API. The differences between these are: REST APIs support connections for short time interval and only limited data can be collected at a time, whereas Streaming API provides tweets in real time and connection for long time. We have used REST API for our analysis. For collecting large amount of tweets, we need long-lived connection and no limit data rate.

To use Twitter API, we must first have a Twitter account. It can be easily created by filling the sign-up details in twitter.com website. After this, you will be provided with a username and password which is used for login purpose. Once the account is created, you can now read and send tweets on any topic you want to explore.

Twitter provides a platform from which we can access data from twitter account and can use it for our own purpose. For this, we have to login with our twitter credentials in dev.twitter.com website. Once our API is created, we can get to know customer key, customer secret key, access token key and access secret key. These keys are used to authenticate user when user wants to access Twitter data.

We created a python script which was used to fetch tweets from Twitter. Before this, we first install a library in Python called tweepy. Tweepy is one of the open-source Python library which enables Python to communicate with Twitter and use its API to collect data so that we can use it in our program.

The script that we used to access data with the help of Twitter is shown in figure 1.

```
import tweepy
from tweepy.streaming import Stream
from tweepy import OAuthHandler
from tweepy import Stream
from tweepy import Cursor
from tweepy import API
import csv
import numpy as np
import pandas as pd
```

```
[ ]  consumer_key = 'Z337WGuWHFncSeOxUqmpelm0S'
     consumer_secret = 'EzG2lvmAEvGVmHY0P2RPiOYS7E3fb6qq2X9lSXoHjR15cvswQx'
     access_token = '920485587286675456-QiSFaR7Zj7iEsDViMKey7RzfBeEXZMY'
     access_token_secret = 'jIFJmz5433DuVIewPFejk24ubN71W01OIw7hHVF9glSSE'
```

```
[ ]  auth = OAuthHandler(consumer_key, consumer_secret)
     auth.set_access_token(access_token, access_token_secret)
```

```
api = tweepy.API(auth,wait_on_rate_limit=True)

csvFile = open('tweet_1.csv', 'a')
#Use csv Writer
csvWriter = csv.writer(csvFile)

for tweet in tweepy.Cursor(api.search_tweets,q="#covid-19",count=100,
                           lang="en",
                           since="2021-02-10",max=10).items():
    print (tweet.created_at, tweet.text)
    csvWriter.writerow([tweet.created_at, tweet.text.encode('utf-8')])
```

Figure 1. Code for getting tweets from Twitter API

The tweets were recorded between 17th February 2022 to 23rd February 2022. We collected around 9K tweets in those seven days.

3.1.2 Data Storage

Once we start getting our data from Twitter API, our next step is to store that data so that we can use it for sentiment analysis. We ran our script for a week and collected tweets. Every time we ran the script, a csv (comma separated values) file is generated which consists of tweets that are extracted from Twitter API. We use .csv file format for our collected data files because data consists of many fields. CSV separate each field with a comma, thus make it easier to access the particular field which consists of text.

3.1.3 Data Pre-Processing

Data obtained from Twitter is not fit for extracting features. Mostly tweets consist of message along with usernames, empty spaces, special characters, stop words, emoticons, abbreviations, hashtags, time stamps, URLs etc. Thus, to make this data fir for mining we pre-process this data by using various functions of NLTK. In pre-processing, we first extract our main message from the tweet and then remove all empty spaces, stop words, hashtags, repeating words, URLs etc. Once we are done with it, we are ready with processed tweet which is provided to classifier for required results. One example is shown below:

Tweet Type	Result
Original tweet	@xyz I think Kejriwal is a habitual liar, even where he don't needs to lie he tells a lie >☹#AAP
Processed tweet	think, habit, lie, even, don't, need, tell, angry

Table 2- Tweets Type

We created code in python in which we define a function which will be used to obtain processed tweet. This code is used to achieve the following functions:

- Remove quotes
- Remove '@'
- Remove URL (Uniform Resource Locator)
- Remove Emoticons: Remove emoticons and replace them with their specific meaning
- Remove duplicates: Remove all repeating words from text so that there will be no duplicates
- Remove '#'
- Remove Stop Words

The table below shows the various types of contents that are included in tweets and also the actions performed on these contents.

CONTENT	ACTION
Punctuation (! ? , . " : ;)	Removed
#word	Removed #word
@any_user	Remove @any_user or replaced with "AT_USER" and then added in stop words.
Uppercase characters	Lowercase all content
URLs and web links	Remove URLs or replaced with "URL" and then added in stop words
Number	Removed
Word not starting with alphabets	Removed
All Word	Stemmed all word (Converted into simple form)
Stop words	Removed
Emoticons	Replaced with respective meaning
White spaces	Removed

Table 3

3.1.4 Tweets Collected

The tweets were recorded between 17th February 2022 to 23rd February 2022. We collected around 9K tweets in those seven days. A sample file for the tweets is shown below:

	A	B
		tweets
0		b'@RexBrynen: Angus Reid poll: most Canadians want protesters to go home (%). Protests have INCREASED support for public health measures\xe\x\xa'
1		b"@NjbBari Do emails in weeks, sent to parents re: Covid from my daughter's school count as data? They're a litt\xe\x\xa
2		b'@nickrmanes: Great historical footnote dug up by @thenighttrain on the Joe Louis Open, a tournament for Black golfers held at Rackham Go\xe\x\xa'
3		b'@NickSawyerMD: COVID- disinformation campaign is the nd Big Lie. \nAmerica\xe\x\xs Frontline Doctors created by the Trump Campaign. Malone h\xe\x\xa'
4		b'@DrEricDing:) There have been more than m excess deaths in the US during the pandemic, according to CDC. The deaths are mainly attrib\xe\x\xa'
5		b"California adopts country's first 'endemic' COVID- plan via @JustTheNews"
6		b'[COVID- Coronavirus outbreak update]\nTotal Cases: „\nTotal Deaths: „\nMore details:\xe\x\xa
7		b'@paimadhu: How did a remote Himalayan district achieve an extraordinary COVID- vaccination rate?\n\n\xe\x\xa@nilanjanab\xe\x\xa captures the amazing w\xe\x\xa'
8		b'@inquirerdotnet: Finance Secretary Carlos Dominguez III proposed on Friday new or higher taxes to pay for the foreign debts the Duterte\xe\x\xa'
9		b'@RebelNewsOnline: Prime Minister Justin Trudeau is facing criticism online after an unsubstantiated claim that his foundation, the Trude\xe\x\xa'
10		b'@VMaya: @DeeEternalOpt DNA vax, Covishield is DNA shred delivered by adenovirus, are known to trigger cancer producing genes. We do\xe\x\xa'
11		b'@DJTTracker: BREAKING: A judge in Massachusetts has blocked the city of Boston from imposing a COVID- vaccination mandate on public sa\xe\x\xa'
12		b'@statsjamie: \xf\xf\xa\xa\xf\xf\xa\xa A further people caught Covid- within welsh hospitals in the w/e //.\n\n\xe\xe\xa\xef\xb\xf It takes the total to , since\xe\x\xa'
13		b'@BNODesk: Biden extends national emergency for COVID-, says the pandemic continues to cause significant risk to public health and safe\xe\x\xa'
14		b'Tricity\xe\x\xs #Covid_ figures for #Saturday\n\n#Chandigarh : cases, death \n#Mohali : cases, death\n#Panchkula:\xe\x\xa
15		b"I admire Novak for his talent, one of the best of all times, but I don't aprove at all his attitude about covid-.\xe\x\xa
16		b'@BharatBiotech: COVAXIN\xe\x\x will be evaluated as a COVID- vaccine candidate in the United States. \n\n#COVID #COVAXIN #VaccineMandate #Bh\xe\x\xa'
17		b'@Parsifaler:) A UNIFYING THEORY OF COVID- PATHOLOGY: THE SPIKE PROTEIN AS SCLEROTHERAPEUTIC\nTRANSFORMING THE VASCULATURE INTO SCAR T\xe\x\xa'
18		b'@OccupyDemocrats: BREAKING: Facebook whistleblower files complaints with the SEC revealing that the social media juggernaut misled inves\xe\x\xa'
19		b'@ImagesByFresh: The GOP are for everything that hurts Americans; they celebrate chaos. \n\nThey are rooting for unemployment, increasing i\xe\x\xa'
20		b'@DJTTracker: BREAKING: A judge in Massachusetts has blocked the city of Boston from imposing a COVID- vaccination mandate on public sa\xe\x\xa'
21		b"@WPR: If I don't want my elementary school-age kids to get COVID-, what's the best way to protect them? What about masks? Do kids need\xe\x\xa"
22		b'There have been concerns that vaccines are less effective against #Omicron. However, mRNA #booster shots offer robu\xe\x\xa
23		b"We Don't Have to Accept a Bad Flu Season Every Winter (#Siguemeytesigo) (#Followback) (#News) (#SDV)"
24		b'@InvestigatorCps: \xf\xf\x\xaAs we move from Pandemic to Endemic, I will be referring to all new COVID- variants as "MAGA FLU."\nAre you listenin\xe\x\xa'

9k

image 1: Sample Tweets Collected

3.1.5 Feature Extraction

Now that we have arrived at our training set, we need to extract useful features from it which can be used in the process of classification. But first we will discuss some text formatting techniques which will aid us in feature extraction:

• Tokenization: It is the process of breaking a stream of text up into words, symbols and other meaningful elements called “tokens”. Tokens can be separated by whitespace characters and/or punctuation characters. It is done so that we can look at tokens as individual components that make up a tweet

• URL’s and user references (identified by tokens “http” and “@”) are removed if we are interested in only analysing the text of the tweet.

• Punctuation marks and digits/numerals may be removed if for example we wish to compare the tweet to a list of English words.

• Lowercase Conversion: Tweet may be normalized by converting it to lowercase which makes it’s comparison with an English dictionary easier.

• Stemming: It is the text normalizing process of reducing a derived word to its root or stem . For example a stemmer would reduce the phrases “stemmer”, “stemmed”, “stemming” to the root word “stem”. Advantage of stemming is that it makes comparison between words simpler, as we do not need to deal with complex grammatical transformations of the word. In our case we employed the algorithm of “**porter stemming**” on both the tweets and the dictionary, whenever there was a need of comparison.

• Stop-words removal: Stop words are class of some extremely common words which hold no additional information when used in a text and are thus claimed to be useless. Examples include “a”, “an”, “the”, “he”, “she”, “by”, “on”, etc. It is sometimes convenient to remove these words because they hold no additional information since they are used almost equally in all classes of text.

A feature is any variable which can help our classifier in differentiating between the different classes. There are two kinds of classification in our system, the objectivity / subjectivity

classification and the positivity / negativity classification. As the name suggests the former is for differentiating between objective and subjective classes while the latter is for differentiating between positive and negative classes.

The list of features explored for positive / negative classification are given below:

• Overall emoticon score (where 1 is added to the score in case of positive emoticon, and 1 is subtracted in case of negative emoticon)

• Overall score from online polarity lexicon MPQA (where presence of strong positive word in the tweet increases the score by 1.0 and the presence of weak negative word would decrease the score by 0.5)

• Number of total emoticons in the tweet

• Number of positive emoticons in a tweet

• Number of negative emoticons in a tweet

• Number of positive words from MPQA lexicon in tweet

• Number of negative words from MPQA lexicon in tweet

• Number of base-form verbs in a tweet

• Number of past tense verbs in a tweet

• Number of plural nouns in a tweet

3.2 Computing Polarity and Subjectivity of Tweets using Textblob

Text Blob is a python library for Natural Language Processing (NLP). Text Blob actively uses Natural Language Toolkit (NLTK) to achieve its tasks. NLTK is a library which gives an easy access to a lot of lexical resources and allows users to work with categorization, classification and many other tasks. Text Blob is a simple library which supports complex analysis and operations on textual data.

Text Blob returns **polarity** and **subjectivity** of a sentence.

Polarity lies between [-1,1], -1 defines a negative sentiment and 1 defines a positive sentiment. Negation words reverse the polarity. Text Blob has semantic labels that help with fine-grained analysis. For example — emoticons, exclamation mark, emojis, etc.

Subjectivity lies between [0,1]. Subjectivity quantifies the amount of personal opinion and factual information contained in the text. The higher subjectivity means that the text contains personal opinion rather than factual information. TextBlob has one more parameter — intensity. TextBlob calculates subjectivity by looking at the '**intensity**'. Intensity determines if a word modifies the next word. For English, adverbs are used as modifiers ('very good').

For example: We calculated polarity and subjectivity for "I do not like this example at all, it is too boring". For this particular example, polarity = -1 and subjectivity is 1, which is fair.

It is expected that if the library returns exactly 0.0 either if your sentence didn't contain any words that had a polarity in the NLTK training set or because TextBlob uses a weighted average sentiment score over all the words in each sample. This easily diffuses out the effect of sentences with widely varying polarities between words in our case: 'helpful' and 'but'.

3.2.1 Working of Textblob

Text blob has a knowledge base in the form of a xml file where for every significant word, except the stop words, polarity values are assigned to each different sense of the word. Same word can be used in different senses.

For e.g. the word "great" has a sense as "very good", or

"Of major significance or importance", or

"Relatively large in size or number or extent", or

"remarkable".Text blob calculates the polarity value for a word by averaging out the polarity values for all the senses of the word. For calculating the polarity value for a word such as "not great" i.e negation of a word, text blob simply multiplies the polarity value for that word with -0.5. There is the concept of intensity for words associated with a modifier. For eg. the word "very great" will have a polarity as 1.0 which is maximum due to the presence of the word

"very" which is a modifier. Thus, the polarity value for "very great" will be calculated by multiplying the polarity of the word "great" with 1.3(for modifier).

A sample for tweets with their respective polarity and subjectivity measures is shown below:

	A	B	C	D	E
		tweets	subjectivity	polarity	sentiment
0		b'@RexBrynen: Angus Reid poll: most Canadians want protesters to go h	0.283333333	0.25	Positive
1		b"@NjbBari Do emails in weeks, sent to parents re: Covid from my daug	0	0	Neutral
2		b'@nickrmanes: Great historical footnote dug up by @thenighttrain on t	0.420833333	0.158333333	Positive
3		b'@NickSawyerMD: COVID- disinformation campaign is the nd Big Lie. \	0.1	0	Neutral
4		b'@DrEricDing:) There have been more than m excess deaths in the US	0.611111111	0.388888889	Positive
5		b"California adopts country's first 'endemic' COVID- plan via @JustTheN	0.333333333	0.25	Positive
6		b'[COVID- Coronavirus outbreak update]\nTotal Cases: „\nTotal Deaths	0	0	Neutral
7		b'@paimadhu: How did a remote Himalayan district achieve an extraorc	0.7	0.277777778	Positive
8		b'@inquirerdotnet: Finance Secretary Carlos Dominguez III proposed on	0.359848485	0.087121212	Positive
9		b'@RebelNewsOnline: Prime Minister Justin Trudeau is facing criticism c	0	0	Neutral
10		b'@VMaya: @DeeEternalOpt DNA vax, Covishield is DNA shred delivere	0	0	Neutral
11		b'@DJTTracker: BREAKING: A judge in Massachusetts has blocked the ci	0.066666667	0	Neutral
12		b'@statsjamie: \xf\xf\xa\xa\xf\xf\xa\xa A further people caught Covid-	0.625	0	Neutral
13		b'@BNODesk: Biden extends national emergency for COVID-, says the p	0.470833333	0.1875	Positive
14		b'Tricity\xe\x\xs #Covid_ figures for #Saturday\n\n#Chandigarh : cases,	0	0	Neutral
15		b"I admire Novak for his talent, one of the best of all times, but I don't a	0.3	1	Positive
16		b'@BharatBiotech: COVAXIN\xe\x\x will be evaluated as a COVID- vacci	0	0	Neutral
17		b'@Parsifaler:) A UNIFYING THEORY OF COVID- PATHOLOGY: THE SPIKE	1	0.5	Positive
18		b'@OccupyDemocrats: BREAKING: Facebook whistleblower files compla	0.333333333	-0.033333333	Negative
19		b'@ImagesByFresh: The GOP are for everything that hurts Americans; th	0	0	Neutral
20		b'@DJTTracker: BREAKING: A judge in Massachusetts has blocked the ci	0.066666667	0	Neutral
21		b"@WPR: If I don't want my elementary school-age kids to get COVID-,	0.6	0.65	Positive
22		b'There have been concerns that vaccines are less effective against #On	0.433333333	0.216666667	Positive
23		b"We Don't Have to Accept a Bad Flu Season Every Winter (#Siguemeyte	0.666666667	-0.7	Negative
24		b'@InvestigatorCps: \xf\xf\x\xaAs we move from Pandemic to Endemic	0.454545455	0.136363636	Positive

9k

Image 2: Sample tweets with their respective polarity and subjectivity values

3.3 Checking Accuracy of Textblob to classify Tweets

Towards the end of the project, our main goal is to study the accuracy and determine which Machine Learning algorithm is performing best in terms of sentiment. The problem is defined as classification, thus we used algorithms such as:

3.3.1 Support Vector Machine (SVM): The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N- the number of features) that distinctly classifies the data points.

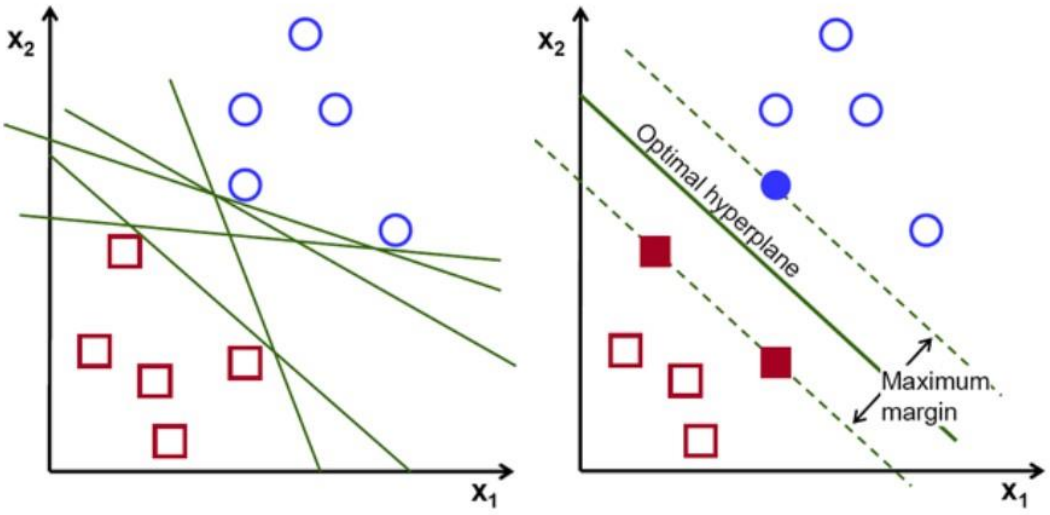


Figure 5 : Possible hyperplanes

To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.

Hyperplanes are decision boundaries that help classify the data points. Data points falling on either side of the hyperplane can be attributed to different classes. Also, the dimension of the hyperplane depends upon the number of features. If the number of input features is 2, then the hyperplane is just a line. If the number of input features is 3, then the hyperplane becomes a two-dimensional plane. It becomes difficult to imagine when the number of features exceeds 3.

3.3.2 Decision Tree : Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

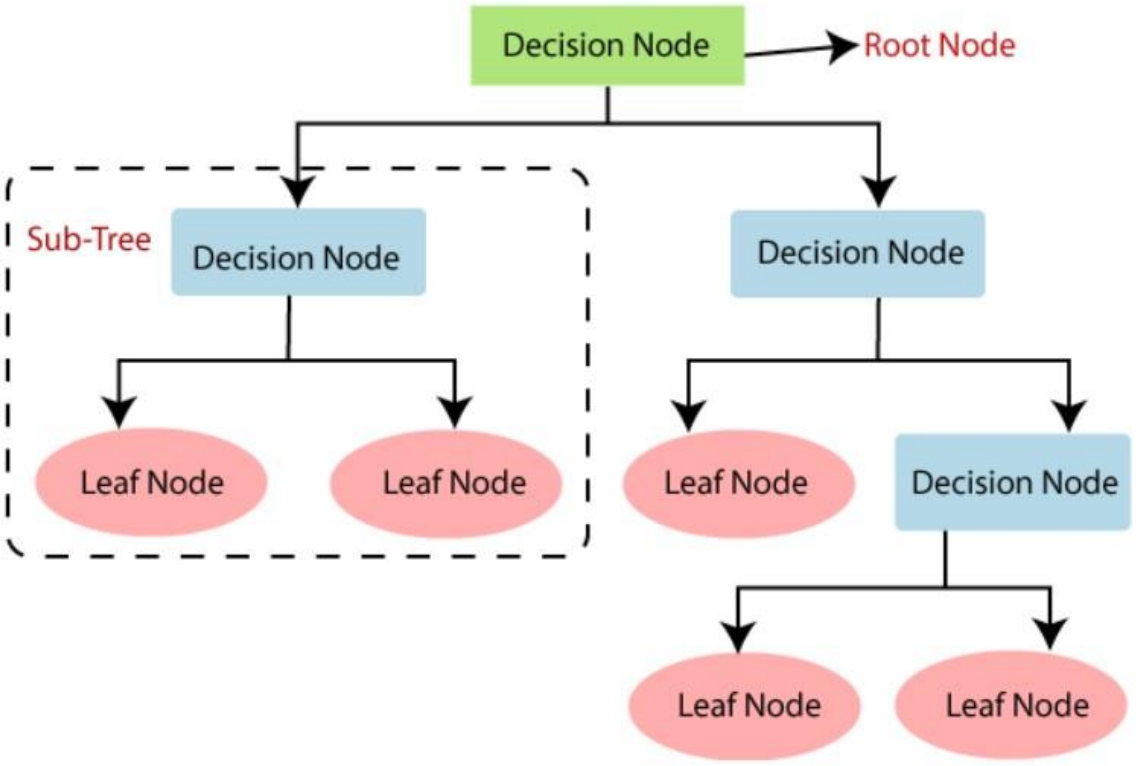


Figure 6: Decision Tree

3.3.3 Logistic Regression: Logistic regression is a statistical method that is used for building machine learning models where the dependent variable is dichotomous: i.e., binary. Logistic regression is used to describe data and the relationship between one dependent variable and one or more independent variables. The independent variables can be nominal, ordinal, or of interval type.

The name "logistic regression" is derived from the concept of the logistic function that it uses. The logistic function is also known as the sigmoid function. The value of this logistic function lies between zero and one.

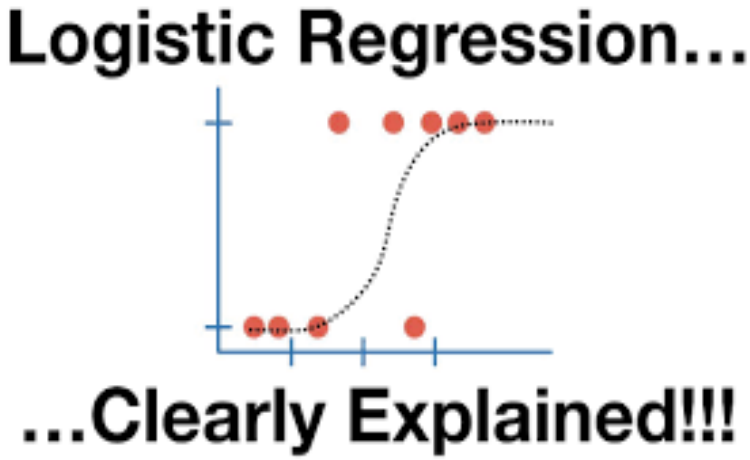


Figure 7: Logistic Regression

3.3.4 SVM- stochastic Gradient Descent: Stochastic gradient descent (SGD) is an algorithm to train the model. According to the documentation, SGD algorithm can be used to train many models.

Gradient Descent is an iterative learning process where an objective function is minimized according to the direction of steepest ascent so that the best coefficients for modeling may be converged upon.

3.3.5 MultinomialNB : Multinomial Naive Bayes classifiers can be used for classification with discrete features (e.g., for text classification). Integer feature counts are normally required for the multinomial distribution.

Multinomial Naive Bayes algorithm is a probabilistic learning method that is mostly used in Natural Language Processing (NLP). The algorithm is based on the Bayes theorem and predicts the tag of a text such as a piece of email or newspaper article.

.

3.3.6 BernoulliNB: This classifier is also suitable for discrete data, just like MultinomialNB. MultinomialNB works with occurrence counts, whereas BernoulliNB is designed to work with binary/boolean features.

The assumption in this model is that the features are binary (0s and 1s) in nature. An application of Bernoulli Naïve Bayes classification is Text classification with 'bag of words' model. The Scikit-learn provides sklearn.

3.3.7 Multi-Level-perceptron: A multilayer perceptron (MLP) is a feed-forward artificial neural network that derives a set of outputs from a set of inputs. As a directed graph, MLPs are composed of several layers of input nodes.
Multi layer perceptron (MLP) is a supplement of feed forward neural network. It consists of three types of layers—the input layer, output layer and hidden layer. The input layer receives the input signal to be processed.

4.RESULT AND CONCLUSION

4.1 RESULT:

In this project, we have analysed the accuracy of the sentiment by using different algorithms. We found that the SVM-SGD is the best, but after the Grid search and other parameters setting, Logistic regression performed better in our case. SVM-SGD is the best for smaller data like.We looked at the tweets and analysed them. The confusion matrix graph depicts how various algorithms are functioning.

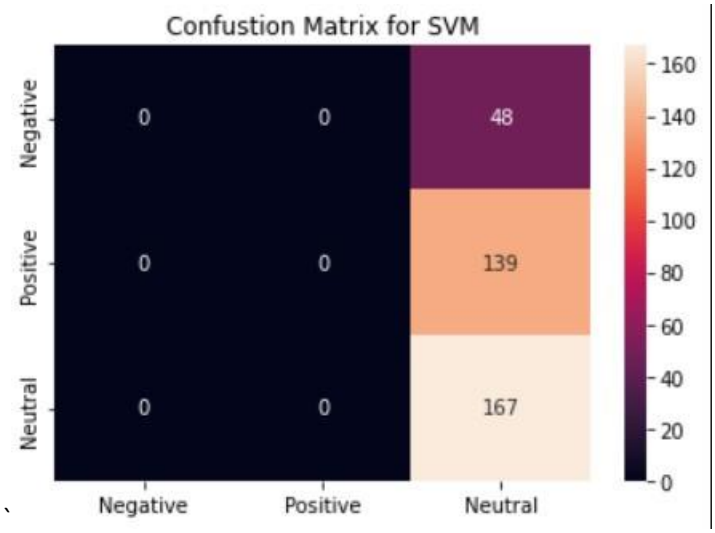


Figure 8: Confusion Matrix for SVM

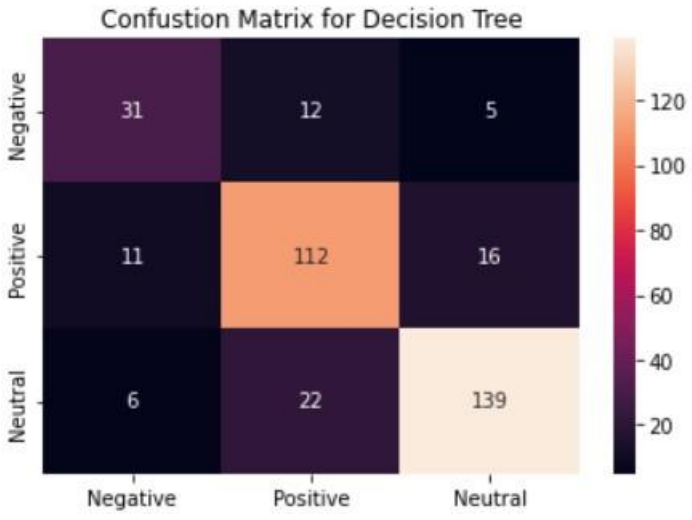


Figure 9: Confusion Matrix for Decision Tree

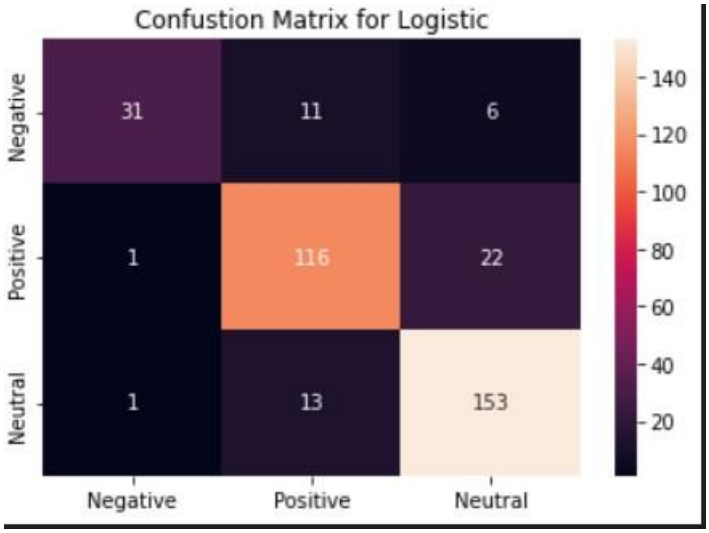


Figure 10: Confusion matrix for Logistic Regression

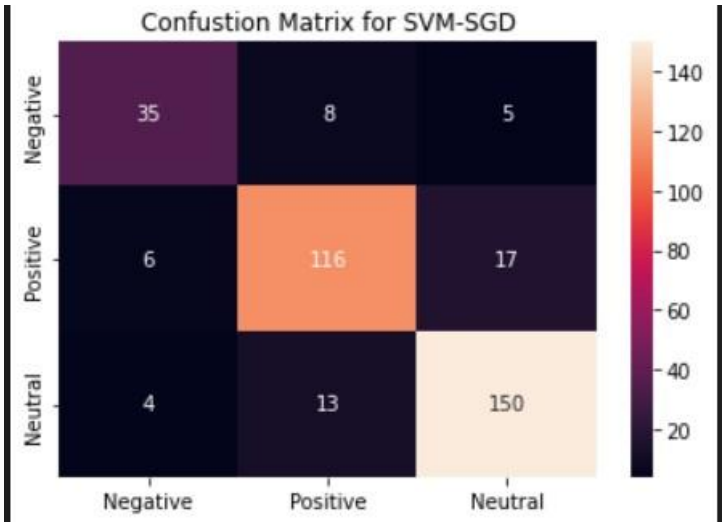


Figure 11: Confusion matrix for SVM-SGD

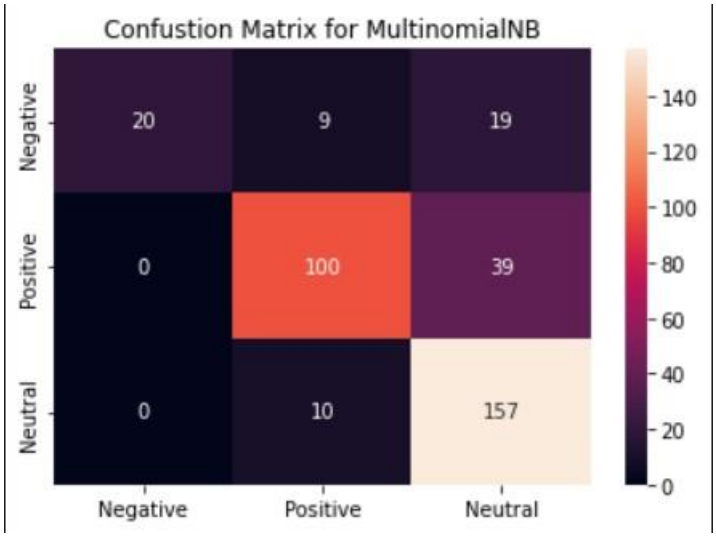


Figure 12: Confusion matrix for MultinomialNB

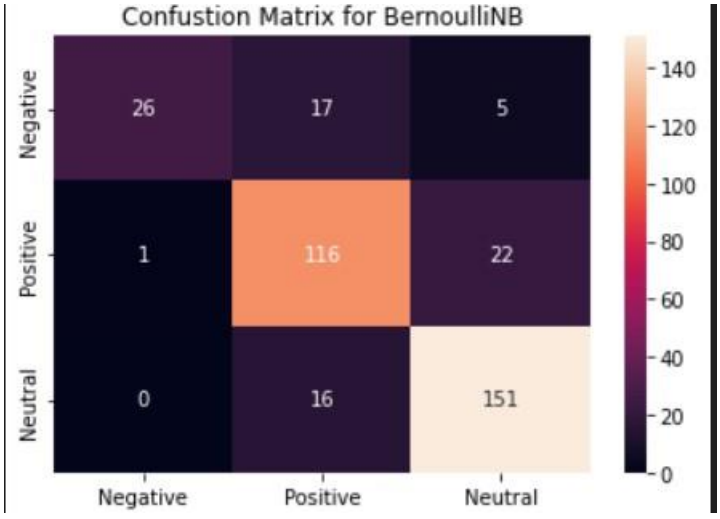


Figure 13: Confusion matrix for BernoulliNB

4.2 FUTURE SCOPE

The future work can be extended to more different algorithms and trying to increase the accuracy by different transformer like Fin sentiment in the Brat , here we are only using Deep Learning we can apply algorithms like LSTM , Bi-LSTM, And building a platform that can analyse any type of the text , not only tweets.

5. REFERENCES

1.Bo Pang and Lillian Lee ,A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts, 2004.

2. Pand and Lee, Sentiment Analysis and Opinion Mining in 2008 and Liu by in 2012.

3.Vishal Shirsath,Rajkumar S Jagdale,Kanchan. V. Shende Sentence Level Sentiment Analysis from News Articles and Blogs using Machine Learning Techniques 2019,pg: 1-2.

4.Lu et al proposed an interactive rule attention network (IRAN) considering the inceuence of grammatical rules,2020,pg 2-3

5.Redex document

6.Stopwords

7.TextBlob

8.Tweepy API

9. Sentiment Analysis using Text Blob

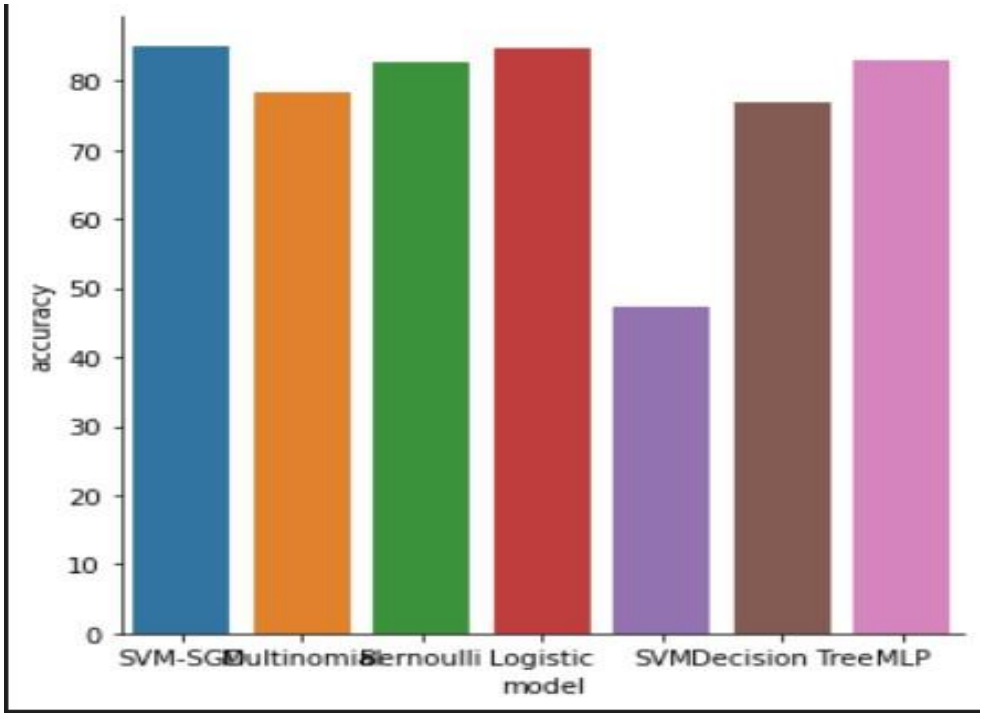
10.MultinomialNB and library usecases

11.BernoulliNB and usecases

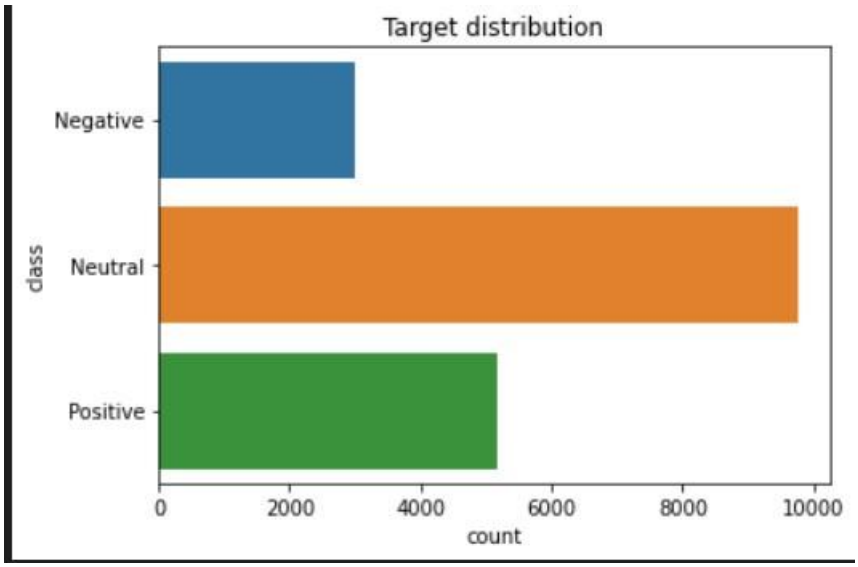
6. APPENDICES

Accuracy of different algorithms:

These different Algorithm show the following results after the train and testing on the data we acquired during the process.



Target distribution



Published term paper