Independent Study

Spring Semester - 2021

SENTIMENT ANALYSIS

SUBMITTED BY:-

Anjali Bhatnagar 2019201012

TASKS UNDER CONSIDERATION

The aim of this project is understanding the sentiment analysis and how we can analyse the sentiment of the data that is available to us.
Before we move on with the details of our problem at hand lets just have a brief overview of the steps that we have followed.

Flow of the Project :

- Data Gathering
- Data Preprocessing
- Training different models for the given data
- Analyzing the predictions of different models and comparing these models
- Comparison of the models
- Selection of the Final Model
- Analysis of some COVID related tweets
- Using the best model to determine the sentiment of a subset of these tweets
- Analysis results

Introduction

- ## What is Sentiment Analysis ?

Sentiment Analysis is the process of determining whether a piece of writing is positive, negative or neutral. With positive sentiment text we mean to convey that the essence with which the sentence is spoken does not do any harm to the sentiments of any person and does not carry a message which contains some form of hatred.

- ## How does Sentiment Analysis work ?

A sentiment analysis system for text analysis combines natural language processing (NLP) and machine learning techniques to assign weighted sentiment scores to the entities, topics, themes and categories within a sentence or phrase. We process the data that is available to us and then using NLP techniques we create the training data. Then we use ML models to train this data which is further used for predicting the sentiment of the data concerned.

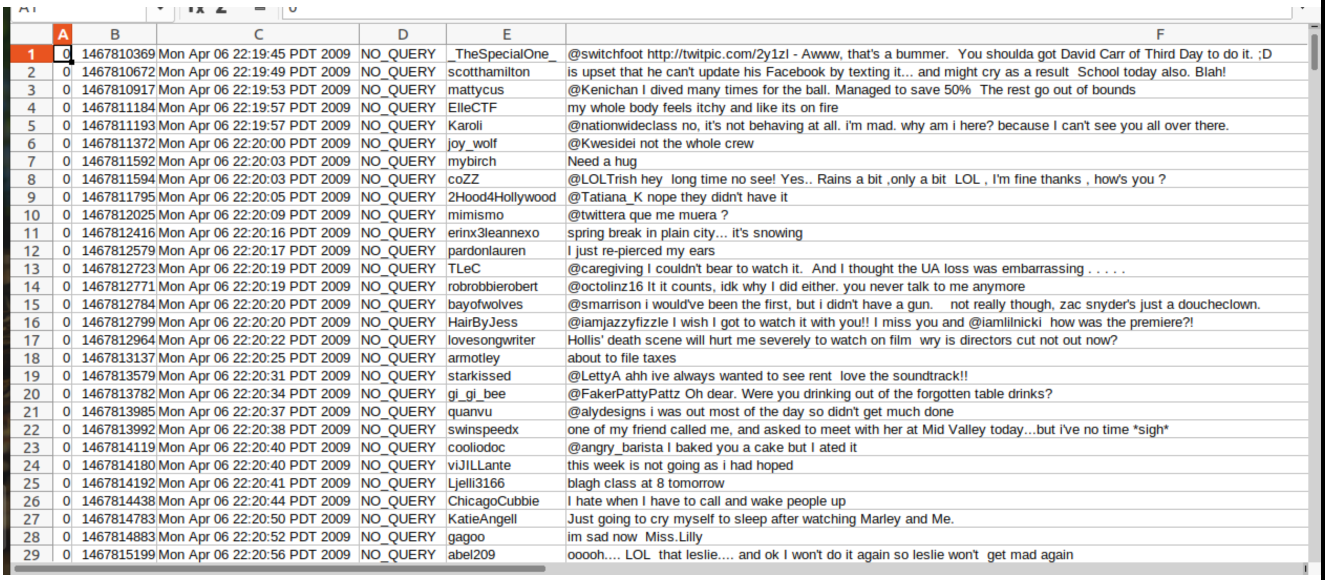
- ## What are the uses of Sentiment Analysis ?

The applications of sentiment analysis are broad and powerful. The ability to extract insights from social data is a practice that is being widely adopted by organisations across the world.

Data Gathering:

The data that we have used for this particular problem of Sentiment Analysis is taken from the Sentiment 140 dataset. It contains 1,600,000 tweets extracted using the twitter api . The tweets have been annotated (0 = negative, 4 = positive) and they can be used to detect sentiment.

This is how the data looks.



	A	B	C	D	E	F
1	0	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zl - Awww, that's a bummer. You shoulda got David Carr of Third Day to do it. ;D
2	0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by texting it... and might cry as a result School today also. Blah!
3	0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Managed to save 50% The rest go out of bounds
4	0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire
5	0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all. i'm mad. why am i here? because I can't see you all over there.
6	0	1467811372	Mon Apr 06 22:20:00 PDT 2009	NO_QUERY	joy_wolf	@Kwesidei not the whole crew
7	0	1467811592	Mon Apr 06 22:20:03 PDT 2009	NO_QUERY	mybirch	Need a hug
8	0	1467811594	Mon Apr 06 22:20:03 PDT 2009	NO_QUERY	coZZ	@LOLTrish hey long time no see! Yes.. Rains a bit ,only a bit LOL , I'm fine thanks , how's you ?
9	0	1467811795	Mon Apr 06 22:20:05 PDT 2009	NO_QUERY	2Hood4Hollywood	@Tatiana_K nope they didn't have it
10	0	1467812025	Mon Apr 06 22:20:09 PDT 2009	NO_QUERY	mimismo	@twittera que me muera ?
11	0	1467812416	Mon Apr 06 22:20:16 PDT 2009	NO_QUERY	erinx3leannexo	spring break in plain city... it's snowing
12	0	1467812579	Mon Apr 06 22:20:17 PDT 2009	NO_QUERY	pardonlauren	I just re-pierced my ears
13	0	1467812723	Mon Apr 06 22:20:19 PDT 2009	NO_QUERY	TLeC	@caregiving I couldn't bear to watch it. And I thought the UA loss was embarrassing . . . . .
14	0	1467812771	Mon Apr 06 22:20:19 PDT 2009	NO_QUERY	robrobbierobert	@octolinz16 It it counts, idk why I did either. you never talk to me anymore
15	0	1467812784	Mon Apr 06 22:20:20 PDT 2009	NO_QUERY	bayofwolves	@smarrison i would've been the first, but i didn't have a gun. not really though, zac snyder's just a doucheclown.
16	0	1467812799	Mon Apr 06 22:20:20 PDT 2009	NO_QUERY	HairByJess	@iamjazzyfizzle I wish I got to watch it with you!! I miss you and @iamlilnicki how was the premiere?!
17	0	1467812964	Mon Apr 06 22:20:22 PDT 2009	NO_QUERY	lovesongwriter	Hollis' death scene will hurt me severely to watch on film wry is directors cut not out now?
18	0	1467813137	Mon Apr 06 22:20:25 PDT 2009	NO_QUERY	armotley	about to file taxes
19	0	1467813579	Mon Apr 06 22:20:31 PDT 2009	NO_QUERY	starkissed	@LettyA ahh ive always wanted to see rent love the soundtrack!!
20	0	1467813782	Mon Apr 06 22:20:34 PDT 2009	NO_QUERY	gi_gi_bee	@FakerPattyPattz Oh dear. Were you drinking out of the forgotten table drinks?
21	0	1467813985	Mon Apr 06 22:20:37 PDT 2009	NO_QUERY	quanvu	@alydesigns i was out most of the day so didn't get much done
22	0	1467813992	Mon Apr 06 22:20:38 PDT 2009	NO_QUERY	swinspeedx	one of my friend called me, and asked to meet with her at Mid Valley today...but i've no time *sigh*
23	0	1467814119	Mon Apr 06 22:20:40 PDT 2009	NO_QUERY	cooliodoc	@angry_barista I baked you a cake but I ated it
24	0	1467814180	Mon Apr 06 22:20:40 PDT 2009	NO_QUERY	viJILLante	this week is not going as i had hoped
25	0	1467814192	Mon Apr 06 22:20:41 PDT 2009	NO_QUERY	Ljelli3166	blagh class at 8 tomorrow
26	0	1467814438	Mon Apr 06 22:20:44 PDT 2009	NO_QUERY	ChicagoCubbie	I hate when I have to call and wake people up
27	0	1467814783	Mon Apr 06 22:20:50 PDT 2009	NO_QUERY	KatieAngell	Just going to cry myself to sleep after watching Marley and Me.
28	0	1467814883	Mon Apr 06 22:20:52 PDT 2009	NO_QUERY	gagoo	im sad now Miss.Lilly
29	0	1467815199	Mon Apr 06 22:20:56 PDT 2009	NO_QUERY	abel209	ooooh.... LOL that leslie.... and ok I won't do it again so leslie won't get mad again

Data Description:

It contains the following 6 fields:

1. target: the polarity of the tweet (*0* = negative, *2* = neutral, *4* = positive)
2. ids: The id of the tweet
3. date: the date of the tweet (*Sat May 16 23:58:44 UTC 2009*)
4. flag: The query . If there is no query, then this value is NO_QUERY.
5. user: the user that tweeted
6. text: the text of the tweet

Data Preprocessing

1. Removing Numeric Data:

If we observe the general pattern then numeric data doesnot contribute to hate speech classification marginal data and can thus be ignored. We can break the tweets into words and check whether the word was numeric or not , if it was numeric then we need not consider it. When I incorporated this approach I was able to see appropriate results in the accuracy and hence this approach is included in my model.

2.Removing tagged users or username mentions:

All the usernames or data which was beginning with @ was removed as mentioning the users does not have any relation with the task under consideration.

3.Handling HashTags (#):

Various approaches were tried to handle the hash tag data. First I tried removing # and then separating the owrds using wordsegment but didn't observe any particular improvement in my prediction model. As each preprocessing comes with a cost and if a particular preprocessing is not bearing expected results we can ignore it. Second I tried removing just the '#' symbol abd then the entire text associated with # , the latter flared well and was perhaps incorporated in the final model.

4.Removing website mentions i.e. links of the type https:

Links to websites itself do not imply any specific criteria which could help us and also increases the processing time. So we donot include any such data.

5.Removing stopwords:

Stopwords are words which are general words and do not convey any special meaning . So we remove this data from the main data .

6.Removing punctuation marks and other signs:

The punctuation marks do not add any specific value and add the same meaning whether used in hate speech or in normal speech so these can be ignored.

7.Removing words smaller than length 2:
These words do not convey a specific meaning and it was observed that these can be ignored.

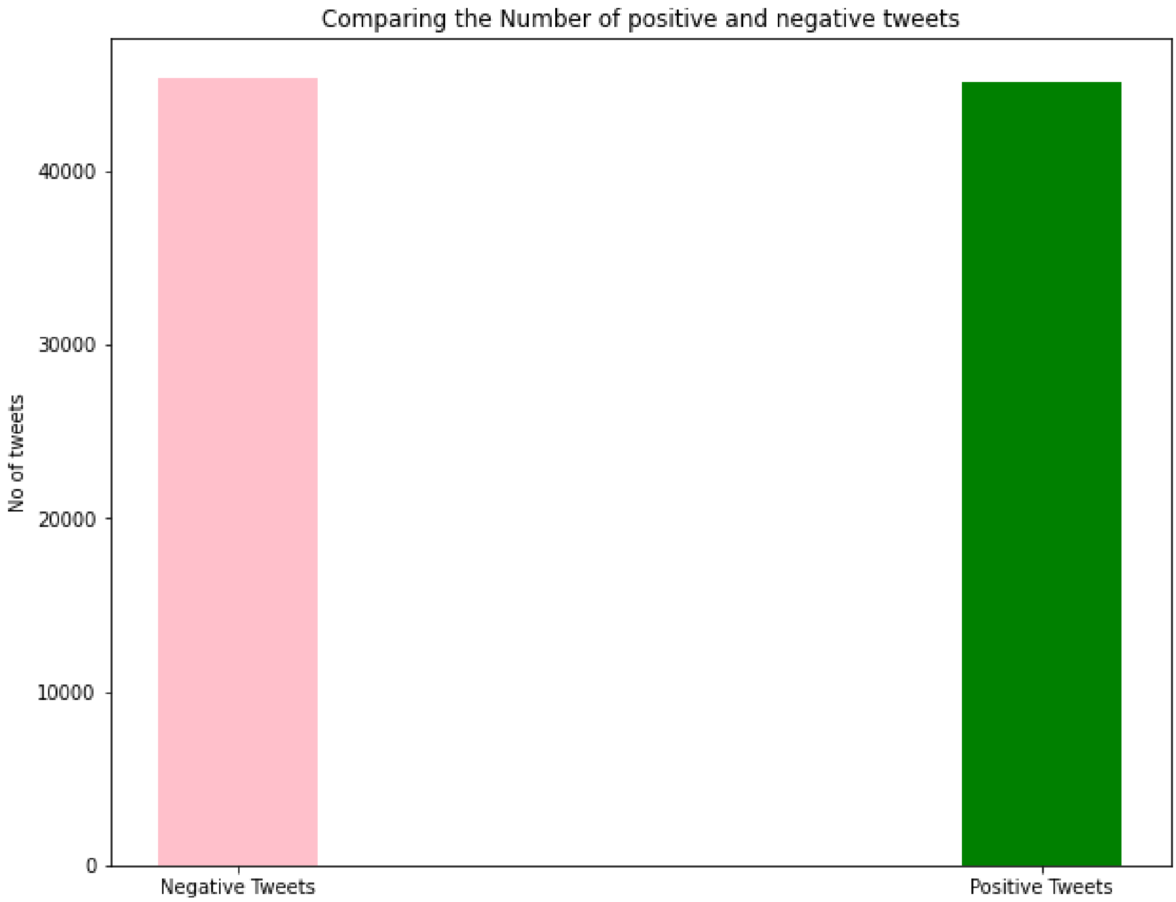
8. Emoticons:
Handling emoticons is a herculean task as it was not possible for a single demoji to decode all the emoticons and then translate them to words and use. Tried it using some words but didnt work out well. Also dropping out emoticons didnt flare out well. So emoticons were left as it is.

9.Using Only Dictionary words:
Using only the dictionary words did not have a huge impact on the predictions but were increasing the processing time considerably so this approach was dropped.

We followed all the above mentioned steps to preprocess the data and to clean the data for further use. The aim of this step was to handle only that data which is important to us and which will help us to classify the sentiment effectively. We went through a lot of experiments to finalise these steps and then concluded with the best data processing approach for the given data.

Data Analysis:

Let us see the data distribution in the dataset. This is done in order to see if there are any biases towards any particular class as this might lead to some incorrect classification results.



From the above analysis it is clear that both the classes contain tweets in equal number so the data is free from any bias and contains an equal representation of both the classes. Now we will start with training this data and we will evaluate the performance of the various models.

Training the Models:

The data after being cleaned and preprocessed was fed to the following 5 models and then evaluated on the basis of the evaluation metrics.

The following are the models that have been used:

- Naive Bayes Classifier
- Logistic Regression
- Decision Tree
- Random Forest Classifier
- Stacking Classifier

Now let us have a look at the results that we obtained from these models and based on this we will be choosing the final model which we will be using on our covid related analysis.

Model 1: Naive Bayes Classifier

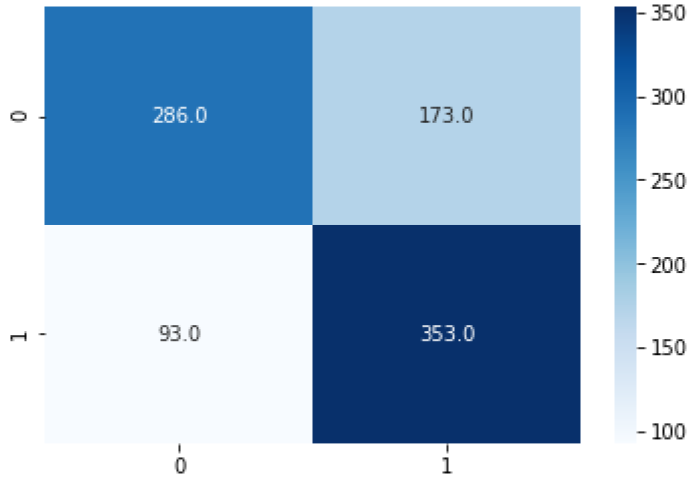
Naive Bayes classifiers are a collection of classification algorithms based on **Bayes' Theorem**. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

When we trained the sentiment data using Naive Bayes Classifier, the results that we got were decent and not any extreme results .

Accuracy : 0.70

F1 Score: 0.71

Confusion Matrix :



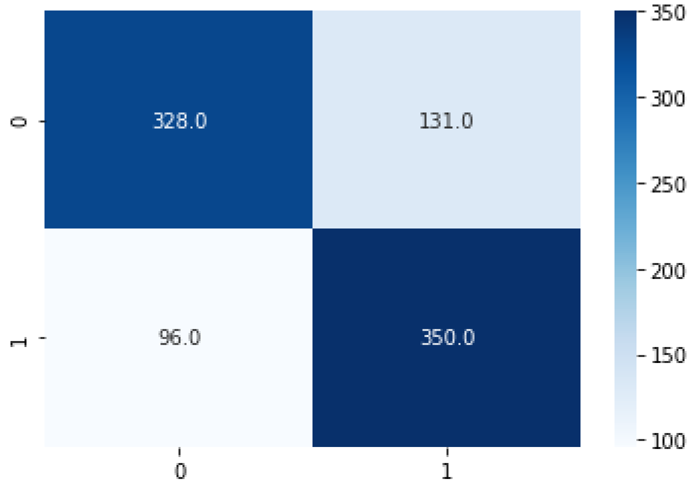
Model 2: Logistic Regression

Logistic regression is a statistical **model** that in its basic form uses a **logistic** function to **model** a binary dependent variable, although many more complex extensions exist. In statistics, the logistic model is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. In our case we use Logistic Regression to decide upon the sentiment of the tweet.

Accuracy : 0.75

F1 Score : 0.74

Confusion Matrix:



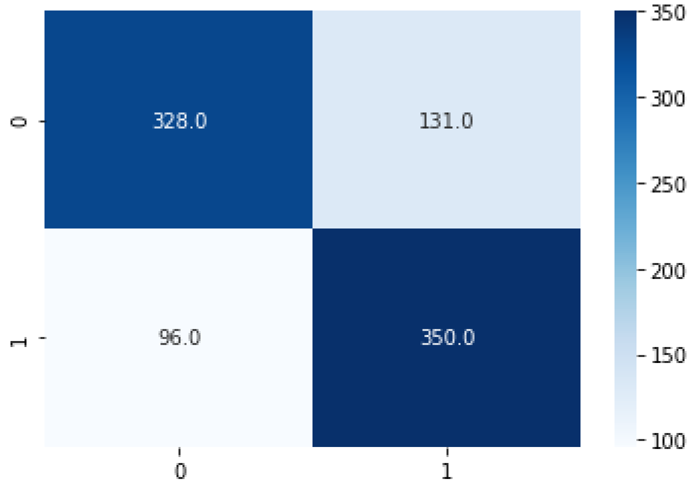
Model 3: Decision Tree

A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements. When we ran the decision tree classifier we got the following results:

Accuracy: 0.69

F1 Score:0.70

Confusion Matrix:



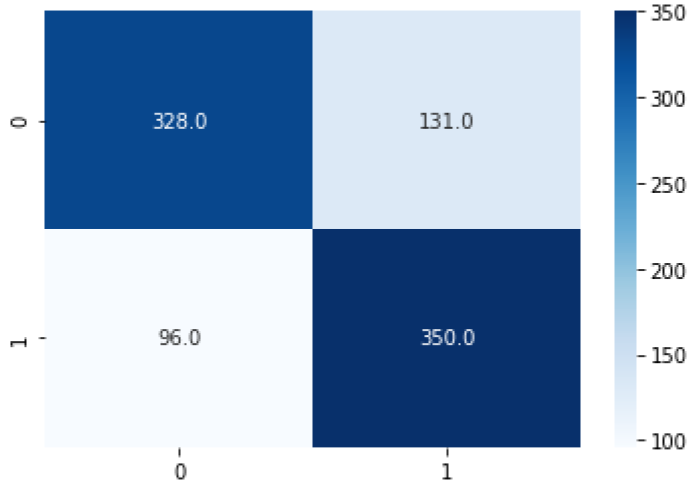
Model 4 : Random Forest Classifier

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. When we ran the random forest classifier we got the following results:

Accuracy: 0.73

F1 Score:0.74

Confusion Matrix:



Model 5: Stacking Classifier

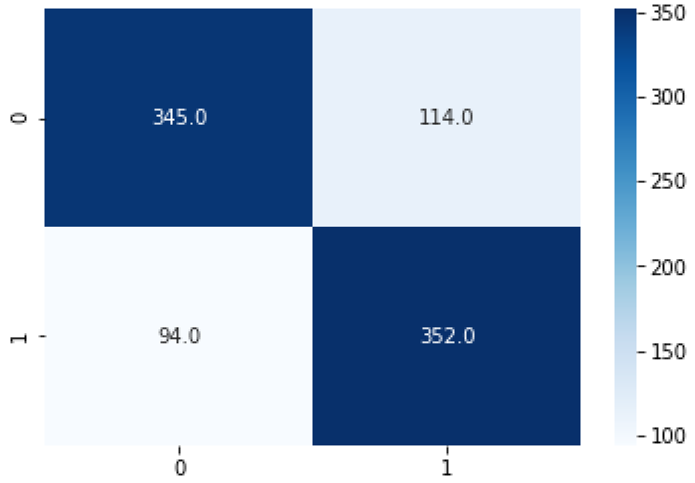
Stacked generalization consists in stacking the output of individual estimator and use a classifier to compute the final prediction. Stacking allows to use the strength of each individual estimator by using their output as input of a final estimator.

In this model we stacked Random Forest Classifier, Naive Bayes Classifier and Logistic Regression classifier together and then used Logistic Regression classifier as the final classifier. We obtained the following results:

Accuracy: 0.78

F1 Score: 0.79

Confusion Matrix:



TASK – COMPARISON BETWEEN THE CLASSIFIERS FOR THIS DATA

For classification purpose we did analysis on our data using the 5 above mentioned models. So before coming on to a conclusion as in which model to use lets have a brief look at the value of the evaluation parameters for all of these classifiers I.e. let us see which model performs best in terms of accuracy and F1 score.

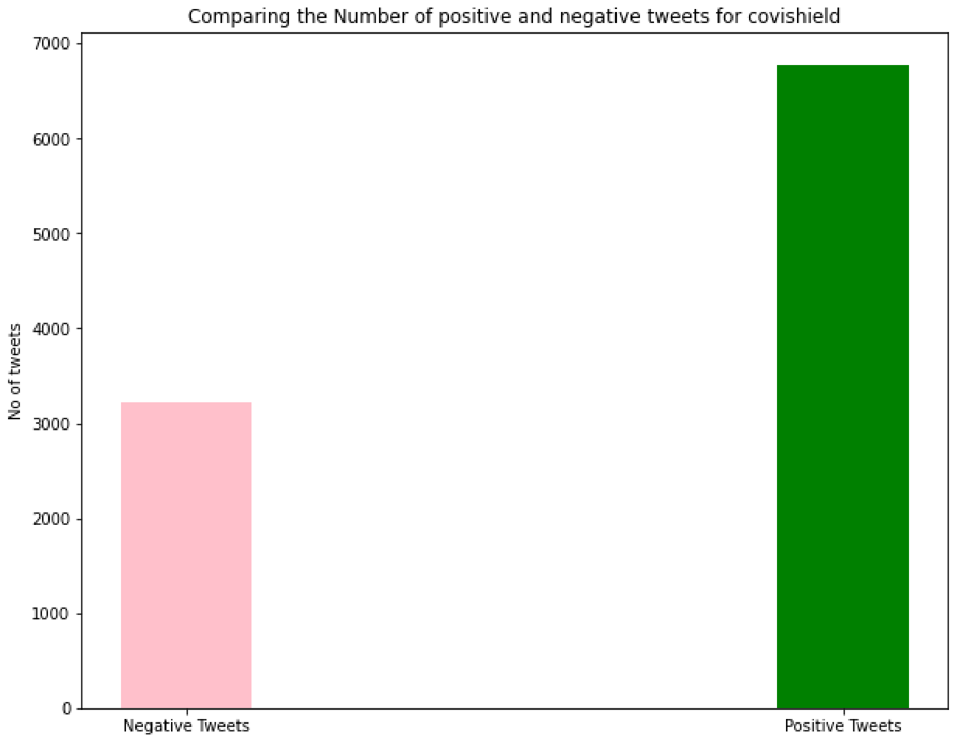
Classifier Name :	**Accuracy Score**	**F1 Score**
Naïve Bayes Classifier	0.70	0.71
Logistic Regression	0.75	0.74
Decision Tree	0.70	0.69
Random Forest	0.73	0.74
Stacking Classifier	0.78	0.79

Thus we can conclude that the Stacking Classifier works best !

Applications of Sentiment Analysis

As an application of sentiment analysis, we will use our best model I.e. the Stacking Classifier to classify some of the Covid related tweets. We trained the model on the data and after we trained it we will be using it to get the sentiments of this data for which the sentiment analysis results are not available to us.

This will help us to see what is the general trend of the tweets related to COVID-19 . I choose to study the tweets on the Covishield Vaccine. I predicted the reults for this data. The results that we obtain are as follows:



Overall we can say that the majority of the tweets about the Covishield Vaccine have a positive sentiment.

Conclusion:

We conclude the sentiment Analysis of the tweets. Further we can use this approach to find the sentiment of many other classes of data which will help us to draw deeper insights about it.

------------------------------THANK YOU--