April 2022

TWITTER SENTIMENT ANALYSIS

A NLP PROJECT

Prepared by:
Sarthak Chandel

What is NLP ?

Natural language processing (NLP) refers to the branch of computer science—and more specifically, the branch of artificial intelligence or AI—concerned with giving computers the ability to understand text and spoken words in much the same way human beings can.

NLP combines computational linguistics—rule-based modeling of human language—with statistical, machine learning, and deep learning models. Together, these technologies enable computers to process human language in the form of text or voice data and to 'understand' its full meaning, complete with the speaker or writer's intent and sentiment.

Sentiment Analysis

Sentiment Analysis, as the name suggests, it means to identify the view or emotion behind a situation. It basically means to analyze and find the emotion or intent behind a piece of text or speech or any mode of communication.

We, humans, communicate with each other in a variety of languages, and any language is just a mediator or a way in which we try to express ourselves. And, whatever we say has a sentiment associated with it. It might be positive or negative or it might be neutral as well.

The goal

The goal was to develop an Automated Machine Learning Sentiment Analysis Model in order to compute the customer perception. Due to the presence of non-useful characters (collectively termed as the noise) along with useful data, it becomes difficult to implement models on them.

Steps taken

In this project, I aimed to analyze the sentiment of the tweets provided from the Sentiment140 dataset by developing a machine learning pipeline involving the use of three classifiers (Logistic Regression, Bernoulli Naive Bayes, and SVM)along with using Term Frequency- Inverse Document Frequency (TF-IDF). The performance of these classifiers is then evaluated using accuracy, F1 Scores and AUC-ROC.

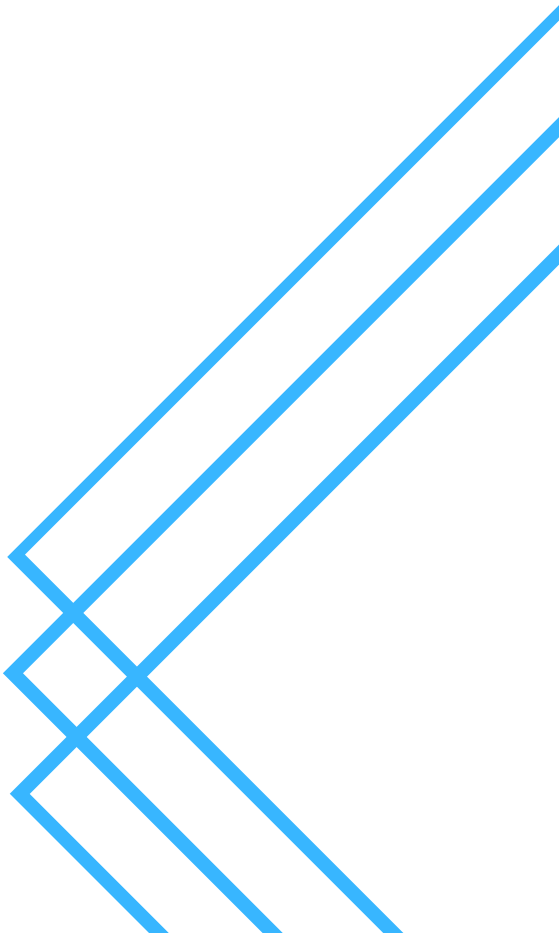
Scope

Sentiment Analysis Dataset Twitter has a number of applications:

Business: Companies use Twitter Sentiment Analysis to develop their business strategies, to assess customers' feelings towards products or brand, how people respond to their campaigns or product launches and also why consumers are not buying certain products.

Politics: In politics Sentiment Analysis Dataset Twitter is used to keep track of political views, to detect consistency and inconsistency between statements and actions at the government level. Sentiment Analysis Dataset Twitter is also used for analyzing election results.

Public Actions: Twitter Sentiment Analysis also is used for monitoring and analyzing social phenomena, for predicting potentially dangerous situations and determining the general mood of the blogosphere.



Dataset Description

The dataset was picked up from kaggle (Sentiment140 Dataset) and it consists of 1,600,000 entries of tweets.

1. target: the polarity of the tweet (0 = negative, 4 = positive)
2. ids: The id of the tweet (2087)
3. date: the date of the tweet (Sat May 16 23:58:44 UTC 2009)
4. flag: The query (lyx). If there is no query, then this value is NO_QUERY.
5. user: the user that tweeted (robotickilldozr)
6. text: the text of the tweet (Lyx is cool)

For our analysis we were only concerned with 2 fields here- the target and the text fields.

Data Preprocessing

Before anything else the first thing we have to do is preprocess our data, so it's clean and ready to be fed into the models with the correct formatting.

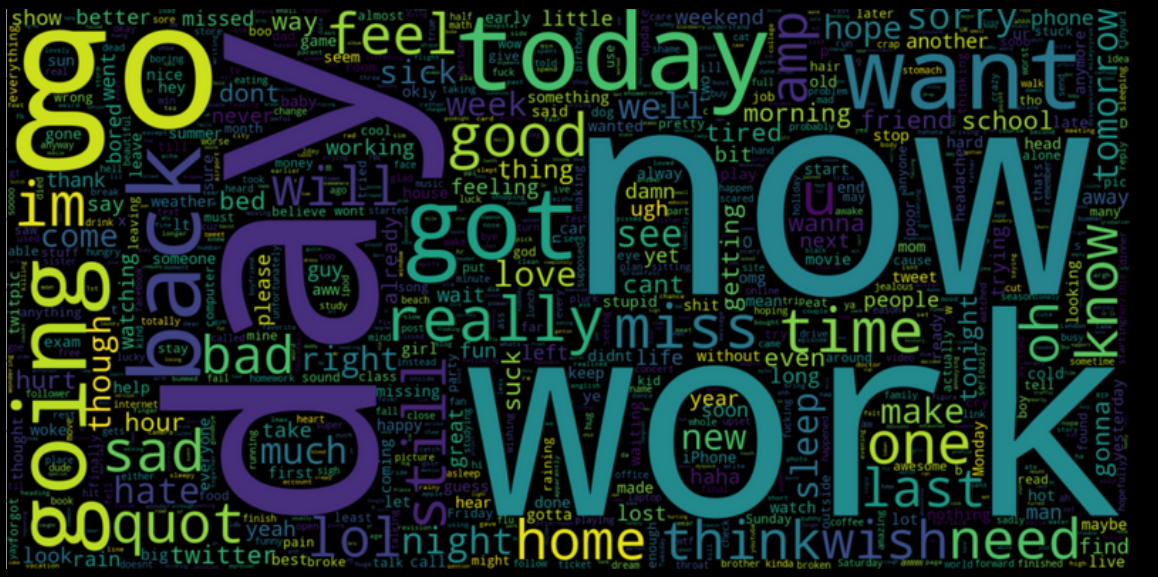
Here are the steps I took to pre-process this data-

1. Converting all text to lower case only
2. Removing stop words from the text. (Stop words are words commonly used in the English language which add no actual value/meaning to the text. eg. a, the, etc.)
3. Getting rid of punctuation.
4. Cleaning repeating characters.
5. Cleaning URLs.
6. Cleaning numerical values.
7. Performing Tokenization. (Tokenization is a way of separating a piece of text into smaller units called tokens.)
8. Performing Stemming and Lemmatization. (Stemming and lemmatization are methods used to analyze the meaning behind a word.)

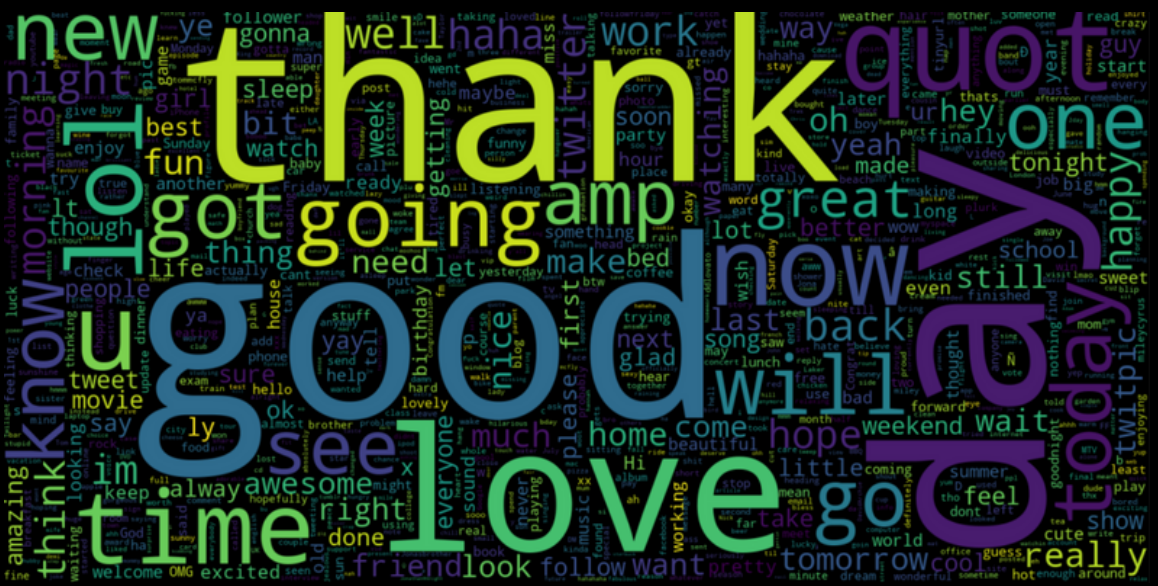
Data Visualization

To better understand and visualize the dataset I was working with, I generated word clouds for both negative and positive tweets seperately.

Negative Tweets WordCloud



Positive Tweets WordCloud



Algorithms Used

I fed the data into 3 different models working on different underlying algos to understand which one worked best.

Bernoulli Naive Bayes

BernoulliNB implements the naive Bayes training and classification algorithms for data that is distributed according to multivariate Bernoulli distributions.

Linear SVC

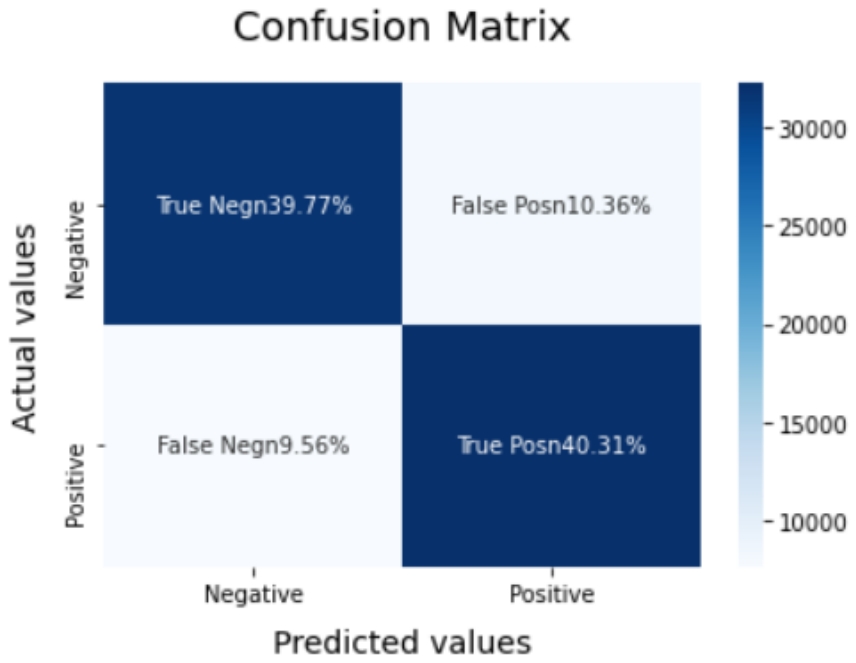
The Linear Support Vector Classifier (SVC) method applies a linear kernel function to perform classification and it performs well with a large number of samples.

Logistic Regression

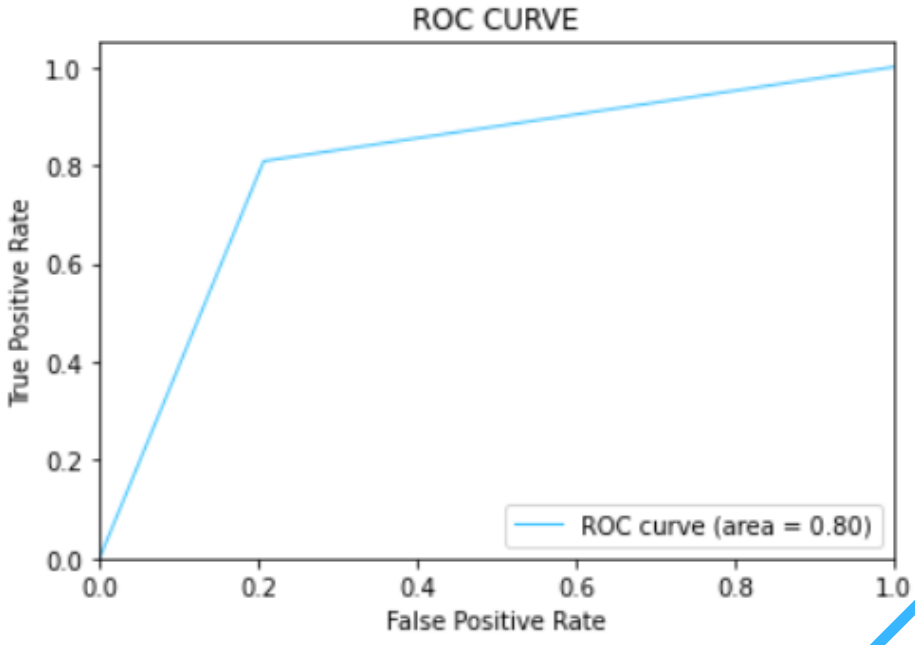
Logistic Regression is a “Supervised machine learning” algorithm that can be used to model the probability of a certain class or event. It is used when the data is linearly separable and the outcome is binary or dichotomous in nature.

Results

Bernoulli Naive Bayes

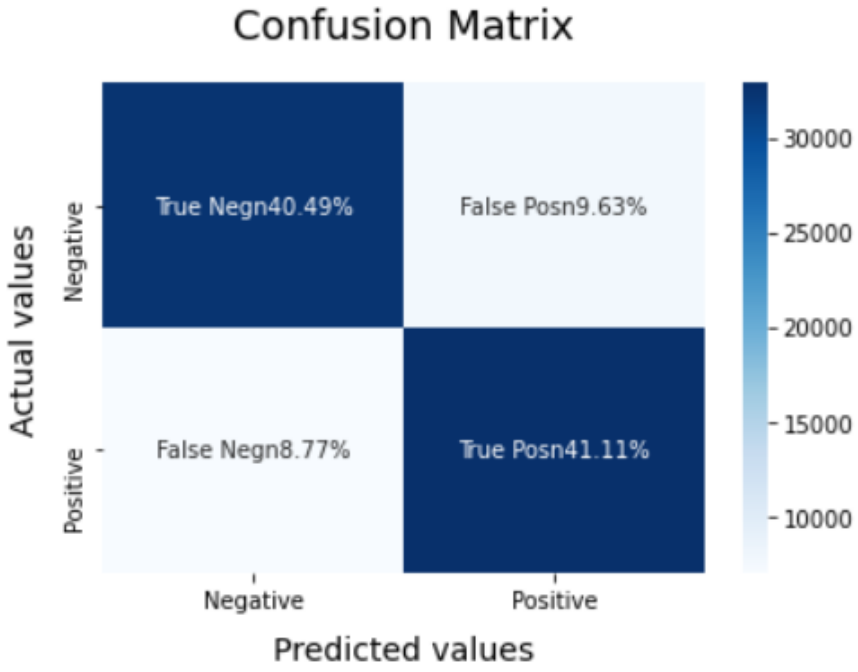


	Precision	Recall	F1-Score
Negative	0.81	0.79	0.80
Positive	0.80	0.81	0.80
Overall	0.80	0.80	0.80

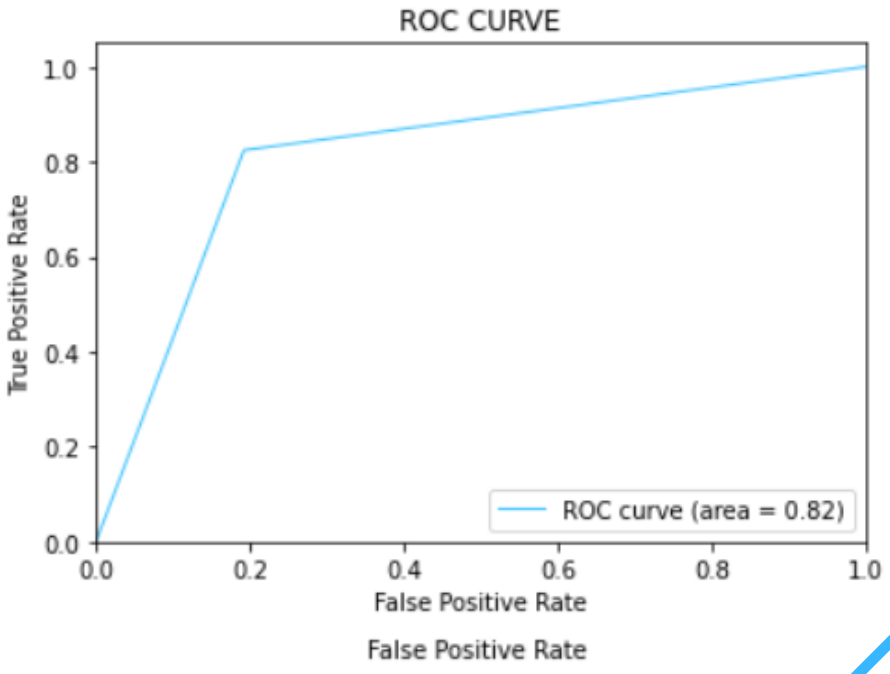


Results

Linear SVC

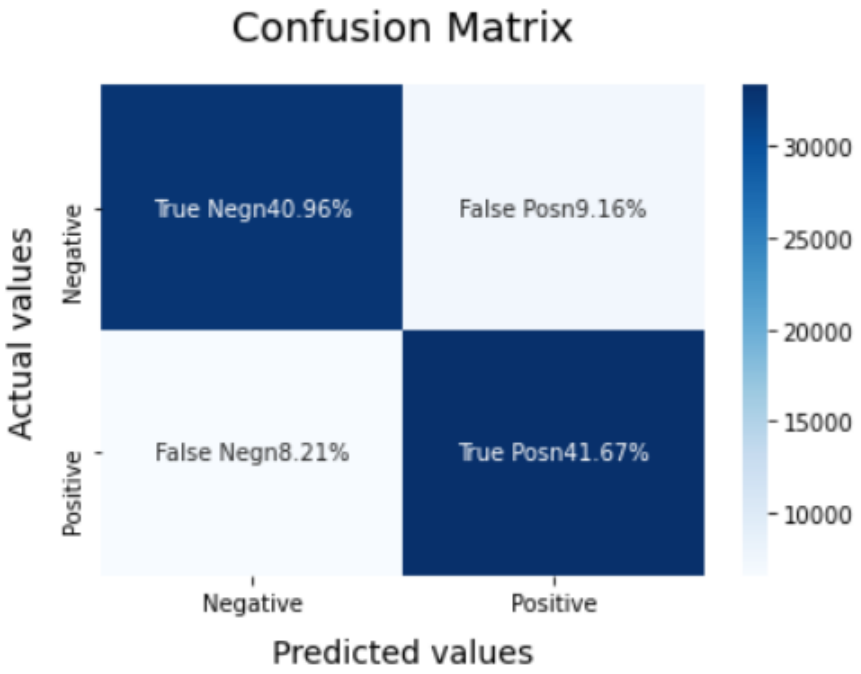


	Precision	Recall	F1-Score
Negative	0.82	0.81	0.81
Positive	0.81	0.82	0.82
Overall	0.82	0.82	0.82

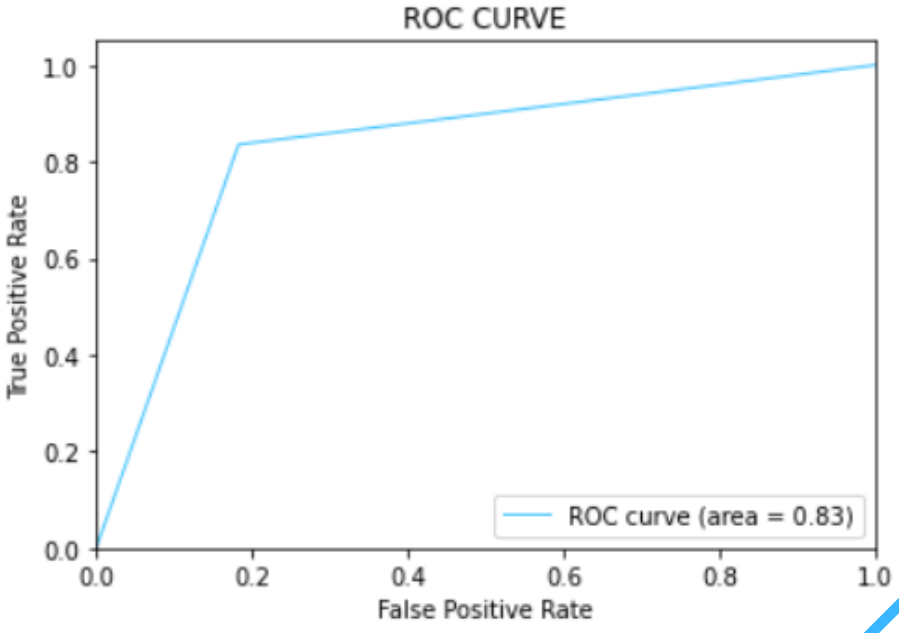


Results

Logistic Regression



	Precision	Recall	F1-Score
Negative	0.83	0.82	0.83
Positive	0.82	0.84	0.83
Overall	0.83	0.83	0.83



Conclusion

Accuracy: As far as the accuracy of the model is concerned Logistic Regression performs better than SVC which in turn performs better than Bernoulli Naive Bayes.

F1-score: The F1 Scores for both classes are :
(a) For Negative class : Bernoulli Naive Bayes(accuracy = 0.80) < SVC (accuracy =0.81) < Logistic Regression (accuracy = 0.92)
(b) For Positive class : Bernoulli Naive Bayes (accuracy = 0.83) < SVC (accuracy = 0.82) < Logistic Regression (accuracy = 0.83)

AUC-ROC: The same observation can be made in terms of ROC area.

Thus, it is safe to assume that for the above given dataset, Logistic Regression outperforms the other models in all parameters and is the best choice to go with.