TWITTER SENTIMENT ANALYSIS

OCT 2020 – NOV 2020

Project by –

ABHISHEK KUBER

SOHAM KULTHE

YASH KULKARNI

QUICK OVERVIEW

Sentiment Analysis is the automated process of analyzing text data and sorting it into sentiments positive, negative, or neutral. Using sentiment analysis tools to analyze opinions in Twitter data can help companies understand how people are talking about their brand.

Twitter boasts 330 million monthly active users, which allows businesses to reach a broad audience and connect with customers without intermediaries. On the downside, there’s so much information that it’s hard for brands to quickly detect negative social mentions that could harm their business.

It can also help understand the political sentiments and viewpoints of people depending upon what they tweet. The trending hashtags on twitter can also be used as the topics of which the sentiments are to be found of.

DATASET

The objective of this task is to detect hate speech in tweets. For the sake of simplicity, we say a tweet contains hate speech if it has a racist or sexist sentiment associated with it. So, the task is to classify racist or sexist tweets from other tweets.

Formally, given a training sample of tweets and labels, where label '1' denotes the tweet is racist/sexist and label '0' denotes the tweet is not racist/sexist, your objective is to predict the labels on the test dataset.

*Dataset taken from Kaggle.*

**IDE :**

Spyder

**COUNT VECTORIZATION FOR TOKENIZATION**

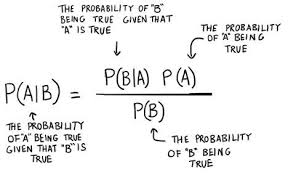
In order to use textual data for predictive modeling, the text must be parsed to remove certain words – this process is called **tokenization**. These words need to then be encoded as integers, or floating-point values, for use as inputs in machine learning algorithms. This process is called **feature extraction (or vectorization)**.

Scikit-learn’s Count vectorizer is used to convert a collection of text documents to a vector of term/token counts. It also enables the ​pre-processing of text data prior to generating the vector representation. This functionality makes it a highly flexible feature representation module for text.

**MACHINE LEARNING MODEL USED:**

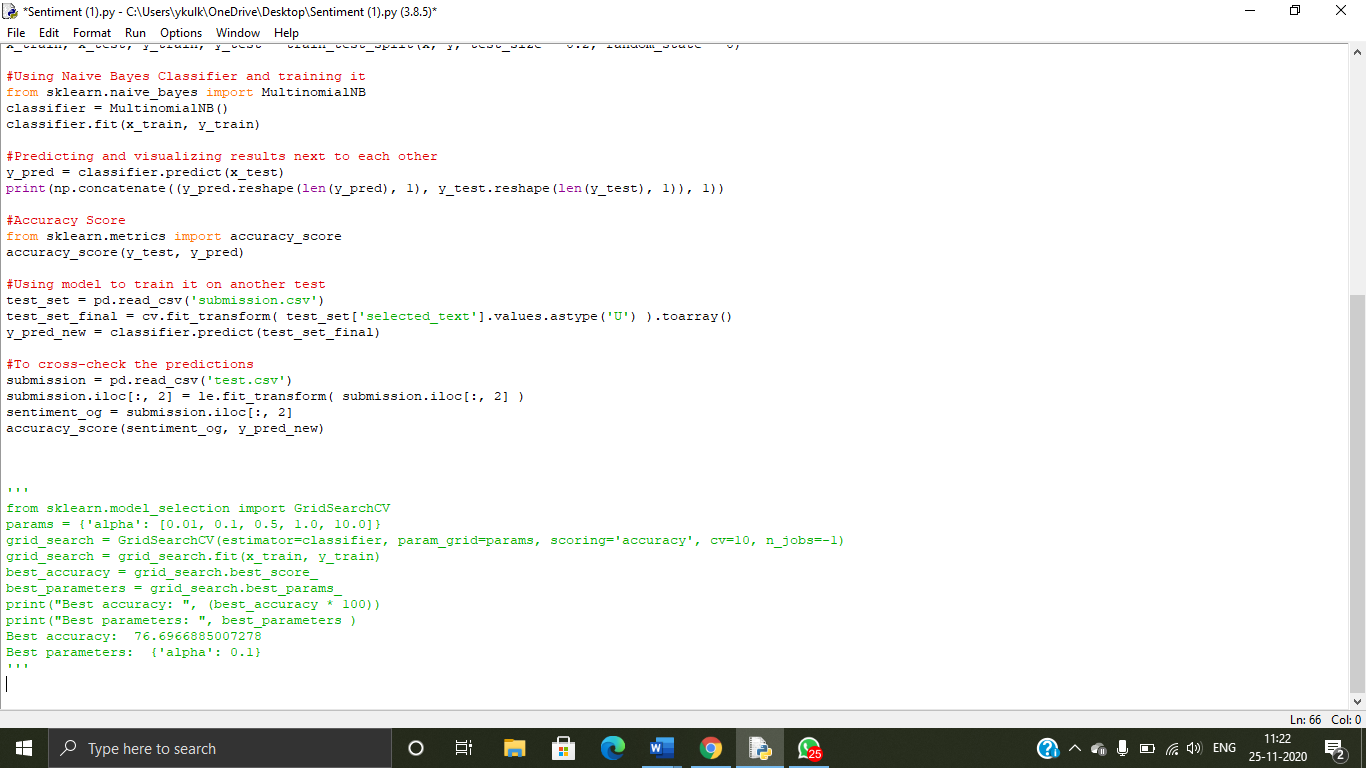
Naïve bayes classifier

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.



**ACCURACY SCORE:**

76.6966885007278



**REFERENCES:**

[*https://www.kaggle.com/arkhoshghalb/twitter-sentiment-analysis-hatred-speech*](https://www.kaggle.com/arkhoshghalb/twitter-sentiment-analysis-hatred-speech)

[*https://en.wikipedia.org/wiki/Naive\_Bayes\_classifier*](https://en.wikipedia.org/wiki/Naive_Bayes_classifier)

*https://scikit-learn.org/stable/modules/generated/sklearn.feature\_extraction.text.CountVectorizer.html*