Dataset

Explaining the dataset has a set of tweets from the users and the data had no null or infinite values the data was already ready for getting trained.

The following dataset was ready to get trained for tweets prediction.

I used a simple classification method to classify the tweets whether the following tweet is from that user or not

Methods used are:

* Count Vectorizer
* TFIDF (Term Frequency Inverse Document Frequency) Vectorizer
* Naïve Bayes
* SVM

Firstly, in model selection, the dataset had a Y value as labeled “author” which can be used to divide for Y\_train and Y \_test.

Then Splitting the data as X\_train and X\_test with a random seed of 53 and with percentage random of 0.33

Feature Extraction:

* Count Vectorizer-

Sklearn provides the inbuilt library for the count vectorizer the main function is to tokenize the given text document and it builds a vocabulary of given words. The following vocabulary makes a new document with the encoded words.

An encoded vector is returned with a length of the entire vocabulary and an integer count for the number of times each word appeared in the document.

* TFIDF (Term Frequency Inverse Document Frequency Vectorizer)

Word count is the best method to categorize the dataset for classification but the problem with the word count is that words like “the” can have a large count so to overcome this issue and also to count the words we use this method term frequency-inverse frequency to count the frequency of the words and then create a vector encoded to create a classification we fit the dataset to naïve Bayes classification.

WE fit the vector then predict the output for the same

The same goes with the Count Vectorizer after creating the encoded vector we fit in the naïve Bayes theorem.

We calculate the accuracy for all the methods and then create the confusion matrix

Overall see in the graph what words are used to after prediction about all the users.