Team members

**شريف أشرف أحمد رشدي**

**عبدالرحمن محمد خليل**

**محمد حسني مسعد**

**يحيى زكريا ابراهيم**

**صبيح صلاح صبيح**

Airline Sentiment Analysis

Natural language processing project

The Dataset

**Twitter US Airline Sentiment Dataset from kaggle:**

**Download from here** (https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment)

**This dataset is about predicting whether the sentiment of the tweets in this set was positive, neutral, or negative for US airlines using some attributes (15 features):**

* Tweet id
* Airline sentiment
* Airline sentiment confidence
* Negative reason
* Negative reason confidence
* Airline
* Airline sentiment gold
* Name
* Negative reason gold
* Retweet count
* Text
* Tweet coord
* Tweet created
* Tweet location
* User time zone

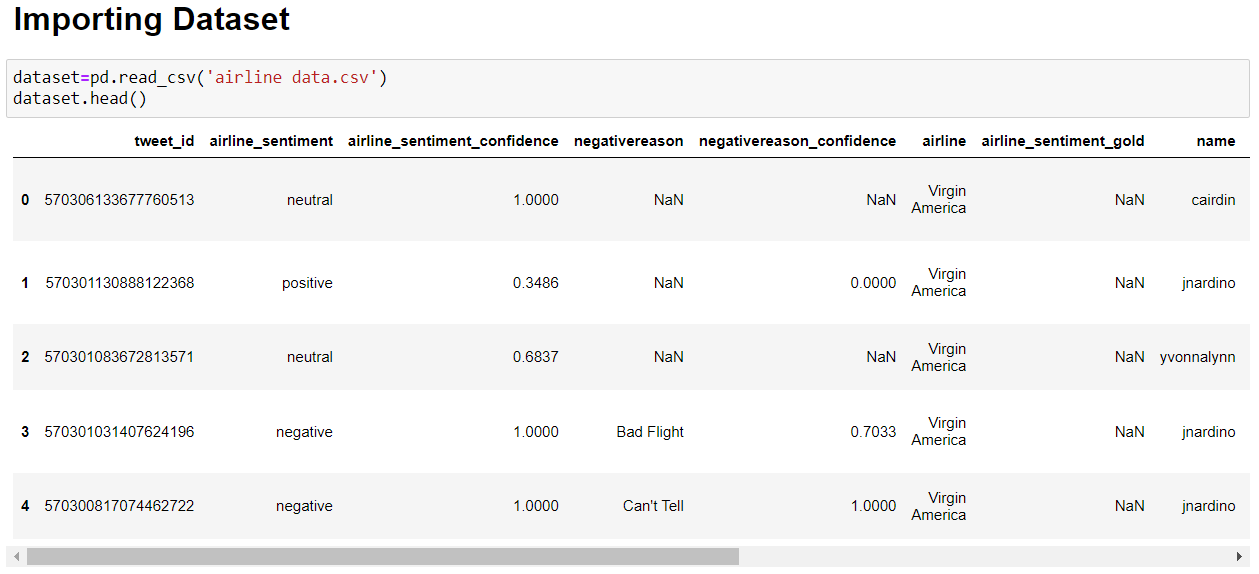
**: The output is**

**Airline Sentiment: positive, neutral, or negative**

**- The dataset file extension is CSV (Comma Separated Values).**

**- So we use Pandas library to read our dataset.**

**- Then we display the first five records with the head function.**

****

**We imported the NLTK library. -**

**- Then we tokenized the original customer reviews from the text column then returning it as list.**

**- Then we made a list documents from the each text and his sentiment label.**



**- In the data preprocessing section:**

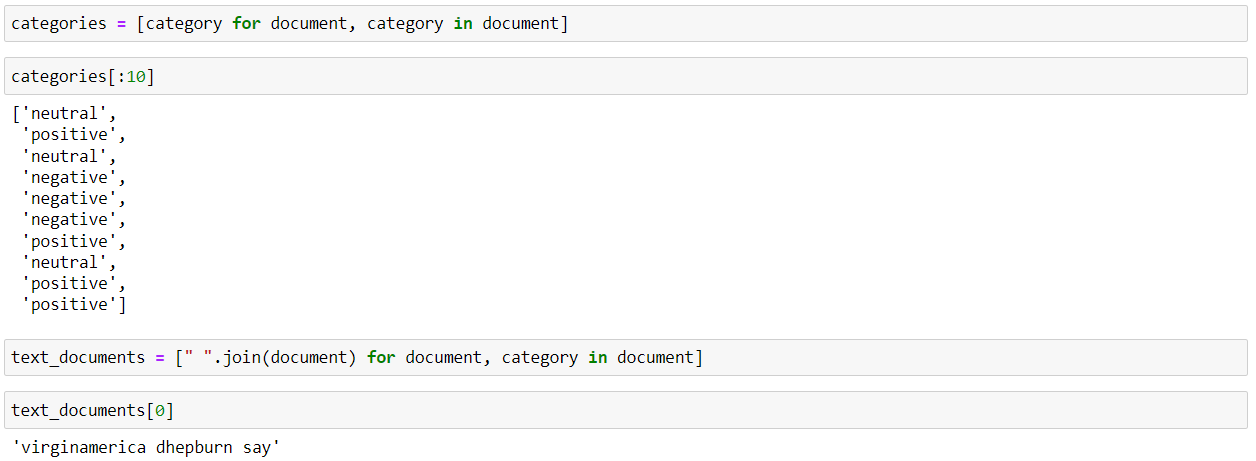
**- We removed the English stop words and the punctuation from the text**

**- We used the lower function to normalize the text**

**- We made the get simple POS function for the POS tagging**

**- We used the Word Net Lemmatizer for the lemmatization**

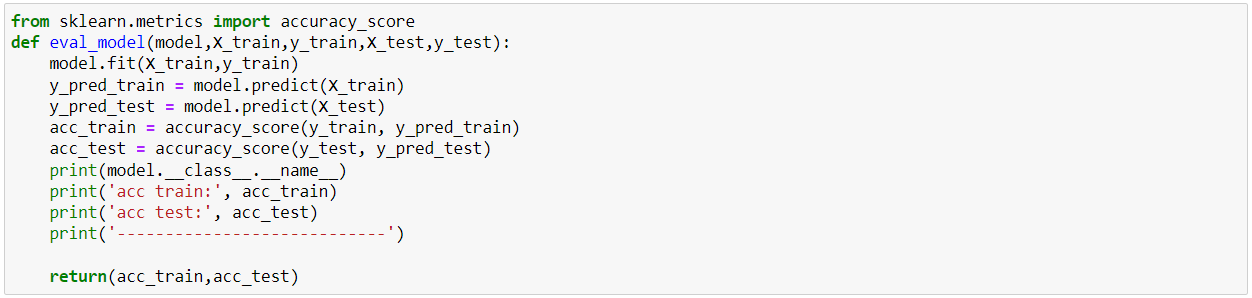
**- Then we disjoined the customer reviews and the sentiment label**

****

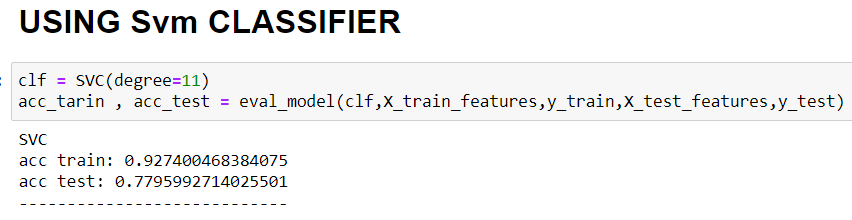
**In the machine learning section: -**

**- We used the train test split function to splitting the dataset into train and test data**

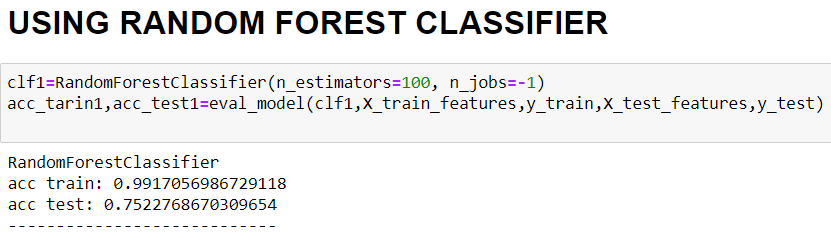
**- We made a helper function to evaluate the model train and test accuracy**



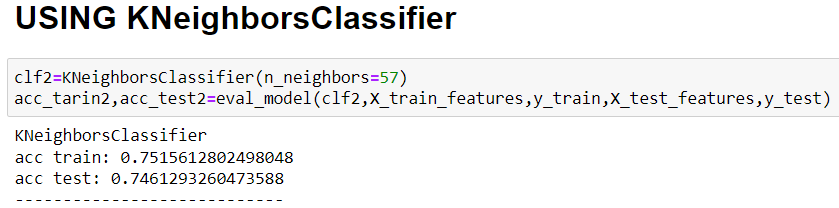
**- When we used the support vector machine classifier it gives 0.92 in the training accuracy and 0.77 in the testing accuracy.**



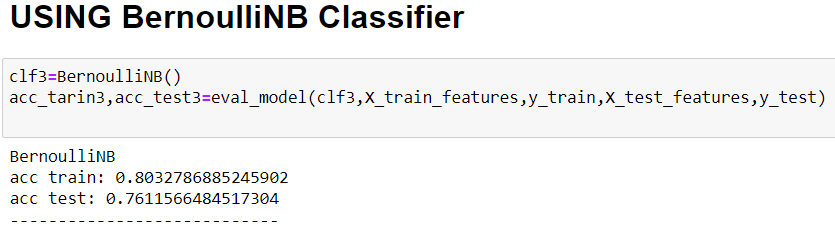
**- When we used the Random Forest Classifier it gives 0.99 in the training accuracy and 0.75 in the testing accuracy.**

****

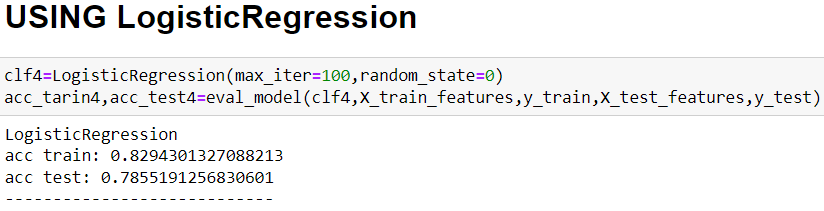
**- When we used the K Neighbors Classifier it gives 0.75 in the training accuracy and 0.74 in the testing accuracy.**

****

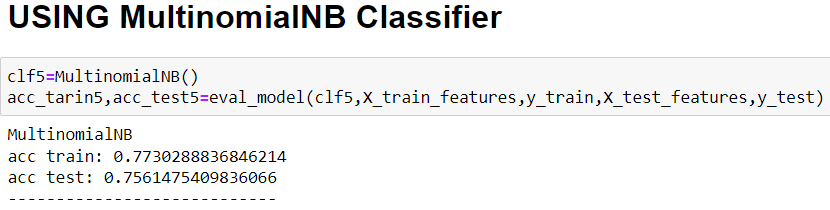
**- When we used the Bernoulli NB Classifier it gives 0.80 in the training accuracy and 0.76 in the testing accuracy.**

****

**- When we used the Logistic Regression Classifier it gives 0.82 in the training accuracy and 0.78 in the testing accuracy.**

****

**- When we used the Multinomial NB Classifier it gives 0.77 in the training accuracy and 0.75 in the testing accuracy.**

****

**We made a voting between the five classification models. -**

**- Then received all the models train and test accuracy in a data frame to display it by the PLT function.**

**- We saved the best model from the voting.**

