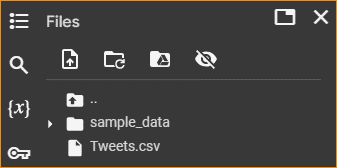
Sentiment analysis is one of the popular downstream applications of Natural Language Processing, which determines the sentiment expressed in a piece of text. The sentiment expressed in a text is usually classified as positive or negative or neutral. You are expected to train a machine learning model to predict the sentiment of Tweets posted about US Airlines.

Step 1 – Load the Dataset

Download the dataset using the following link:

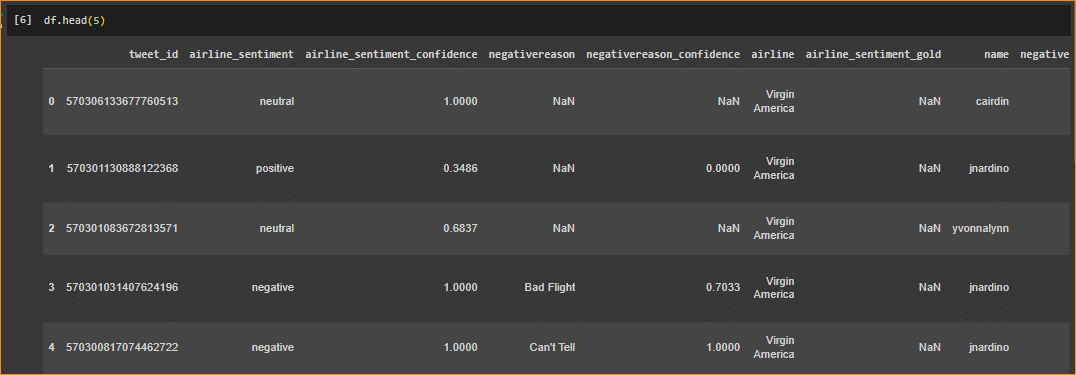
<https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment>

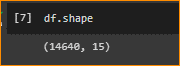
Upload the dataset to Google Colab directly or via Google Drive.



Read the dataset using pandas library.

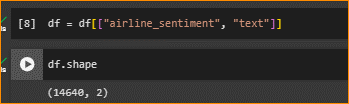


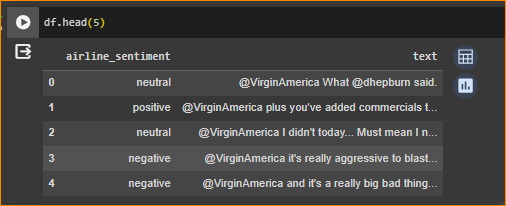




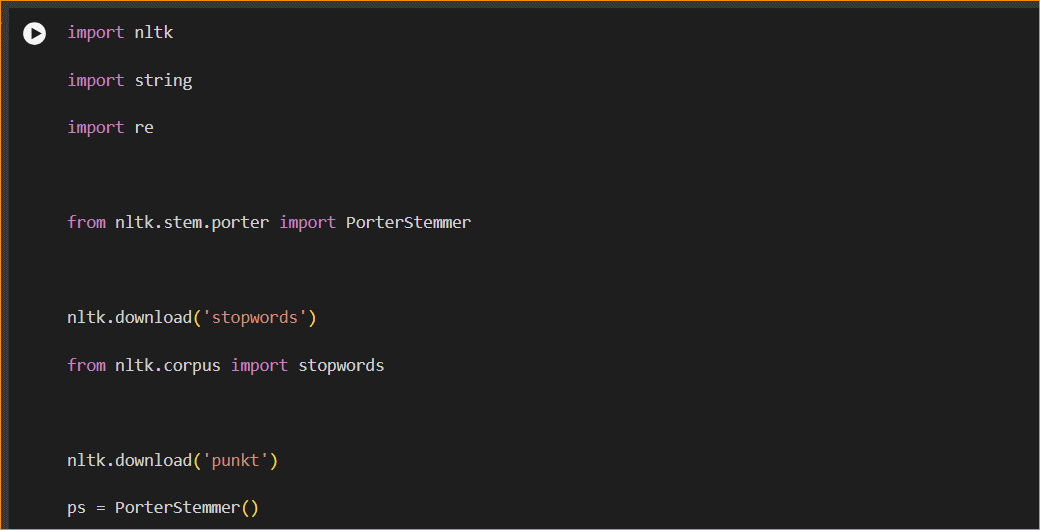
The column “text” contains the tweets written by the users about different Airlines and the column “airline\_sentiment” contains the class label indicating the sentiment of the “text” as positive or negative or neutral. We require only the columns "airline\_sentiment " and " text " for the sentiment analysis task. Hence, drop the other columns in the dataset using the following code snippet.

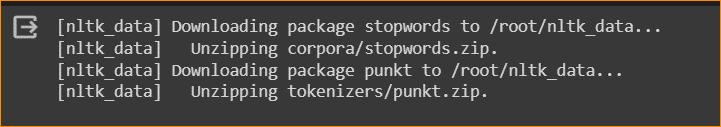
df = df[["airline\_sentiment", "text"]]

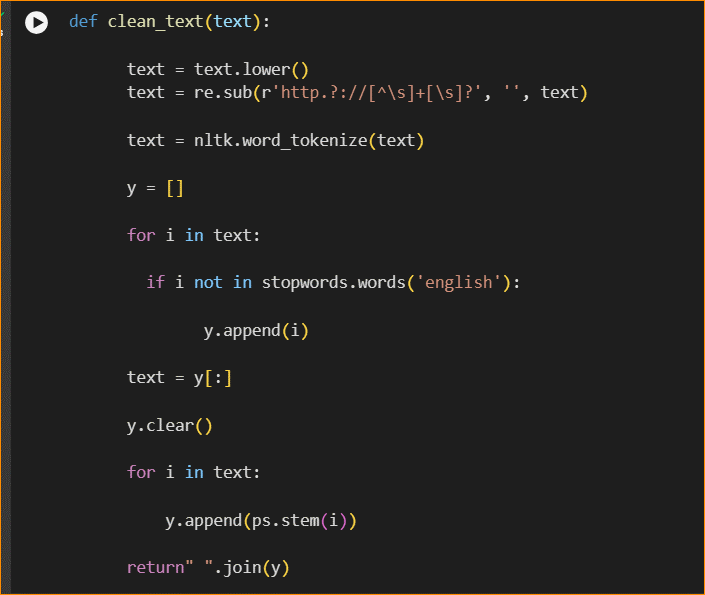




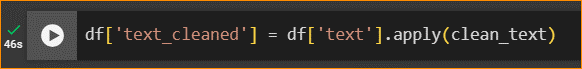
Step 2 – Preprocess Text

The following code snippet preprocesses the text by obtaining lower case, removing URLs, removing stop words, and generating stem. Add the code snippet to your notebook.  




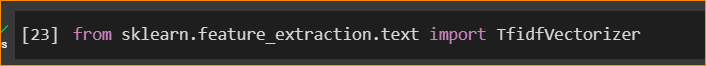


Apply “clean\_text” function to “text” column and assign the resulting text to new column named “text\_cleaned”.

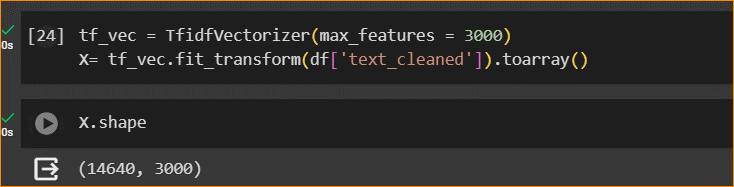


Step 3 – Feature Extraction

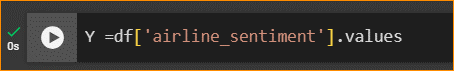
* Create a “TfidfVectorizer” using the “sklearn” library by specifying “max\_features” as 3000



* Use “TfidfVectorizer” created to generate TF-IDF vector representation of cleaned text in the column “text\_cleaned”, convert the vector representation to an array, and assigned it to a new variable named “X”.

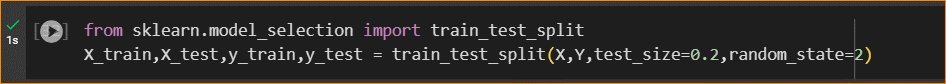


* Convert the column “airline\_sentiment” to an array and assigned it to a new variable named “Y”.

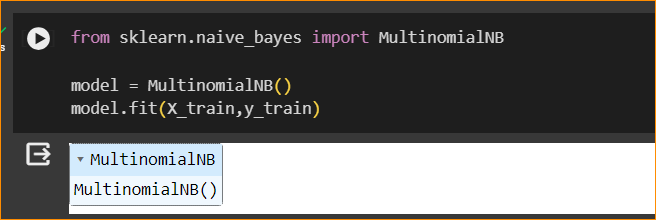


Step 4 – Train Model

* Split the dataset into training and testing sets using the function “train\_test\_split” available in the “sklearn” library. Set the parameters of the functions “test\_size” to 0.2 and “random\_state” to 2.



* Train a multinomial Naïve Bayes classifier using the training dataset and predict the sentiment of tweets in the test dataset. Find the accuracy of the model in predicting sentiment labels for the test dataset.



* Train a Random Forest classifier using the training dataset, predict the sentiment of tweets in the test dataset, and find the accuracy of the model using the following code snippet,

