# **Author: Albright Raj**

# Project Title: Sentiment analysis for **marketing**

# **Dataset Link:**

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

import nltk  
import pandas as pd  
import numpy as np  
from nltk.corpus import stopwords  
from nltk.tokenize import word\_tokenize

nltk.download('stopwords')  
nltk.download('punkt')

[nltk\_data] Downloading package stopwords to /root/nltk\_data...  
[nltk\_data] Package stopwords is already up-to-date!  
[nltk\_data] Downloading package punkt to /root/nltk\_data...  
[nltk\_data] Package punkt is already up-to-date!

True

# **Data Collection**

We need to load the Twitter Airlines Sentiment CSV file into jupyter notebook.

data = pd.read\_csv('Tweets.csv')

data.columns

Index(['tweet\_id', 'airline\_sentiment', 'airline\_sentiment\_confidence',  
 'negativereason', 'negativereason\_confidence', 'airline',  
 'airline\_sentiment\_gold', 'name', 'negativereason\_gold',  
 'retweet\_count', 'text', 'tweet\_coord', 'tweet\_created',  
 'tweet\_location', 'user\_timezone'],  
 dtype='object')

data.head()

tweet\_id airline\_sentiment airline\_sentiment\_confidence \  
0 570306133677760513 neutral 1.0000   
1 570301130888122368 positive 0.3486   
2 570301083672813571 neutral 0.6837   
3 570301031407624196 negative 1.0000   
4 570300817074462722 negative 1.0000   
  
 negativereason negativereason\_confidence airline \  
0 NaN NaN Virgin America   
1 NaN 0.0000 Virgin America   
2 NaN NaN Virgin America   
3 Bad Flight 0.7033 Virgin America   
4 Can't Tell 1.0000 Virgin America   
  
 airline\_sentiment\_gold name negativereason\_gold retweet\_count \  
0 NaN cairdin NaN 0   
1 NaN jnardino NaN 0   
2 NaN yvonnalynn NaN 0   
3 NaN jnardino NaN 0   
4 NaN jnardino NaN 0   
  
 text tweet\_coord \  
0 @VirginAmerica What @dhepburn said. NaN   
1 @VirginAmerica plus you've added commercials t... NaN   
2 @VirginAmerica I didn't today... Must mean I n... NaN   
3 @VirginAmerica it's really aggressive to blast... NaN   
4 @VirginAmerica and it's a really big bad thing... NaN   
  
 tweet\_created tweet\_location user\_timezone   
0 2015-02-24 11:35:52 -0800 NaN Eastern Time (US & Canada)   
1 2015-02-24 11:15:59 -0800 NaN Pacific Time (US & Canada)   
2 2015-02-24 11:15:48 -0800 Lets Play Central Time (US & Canada)   
3 2015-02-24 11:15:36 -0800 NaN Pacific Time (US & Canada)   
4 2015-02-24 11:14:45 -0800 NaN Pacific Time (US & Canada)

# **Data Preprocessing**

This below provided code snippet utilizes the NLTK library to preprocess text data stored in a DataFrame called 'data.' It first initializes a set of English stopwords and then defines a custom function, 'preprocess\_text,' which tokenizes, converts to lowercase, and removes non-alphanumeric characters from each text entry while filtering out stopwords. This function is applied to the 'text' column of the DataFrame, and the cleaned text is stored in a new 'cleaned\_text' column, making the data more suitable for various natural language processing tasks, such as sentiment analysis or text classification.

stop\_words = set(stopwords.words('english'))

def preprocess\_text(text):  
 words = word\_tokenize(text)  
 words = [word.lower() for word in words if word.isalnum() and word.lower() not in stop\_words]  
 return ' '.join(words)  
  
data['cleaned\_text'] = data['text'].apply(preprocess\_text)

X = data['cleaned\_text']  
y = data['airline\_sentiment']

data['cleaned\_text'].head()

0 virginamerica dhepburn said  
1 virginamerica plus added commercials experienc...  
2 virginamerica today must mean need take anothe...  
3 virginamerica really aggressive blast obnoxiou...  
4 virginamerica really big bad thing  
Name: cleaned\_text, dtype: object