In Class Project

Airline Sentiment Analysis

edureka!

In Class Project

The dataset has been taken from Kaggle.
(https://www.kaggle.com/crowdflower/twitter-airline-sentiment/home)

This is a dataset having tweets about 6 US Airlines along with their sentiments: positive, negative and neutral.

You are provided with two files: 'Tweets-train.csv" and 'Tweets-test.csv"

The train data contains about 11000 tweets and test contains 4000 tweets. You have to perform Sentiment Analysis on the dataset and also built a classifier on the training data.

Following are the operations to be carried out:

- Read the training using pandas module and select only the sentiment and text columns

- Observe randomly generated 10 tweets for each sentiment with respect to the following:
    - Text contains references with '@'
    - Text contains links (http , https)
    - Text contains punctuations
    - Text contains Emoticons

- You have to prepare a function to clean all the above observed tokens from the tweet text.
  Save changes in a new column

- List down the most common 15 words for each sentiment. Observe the results

- Remove Stopwords from all the tweets.
  Save changes in a new column and list down most common 15 words.

- Remove these words from all the tweets.

americanair, united, delta, southwestair, jetblue, virginamerica, usairways, flight, plane

Save changes in a new column and list down most common 15 words. Comment your observations

- Encode Sentiments using Label Encoder
- Vectorize the Text Column (You can choose any vectorizer of your choice)
- Prepare a multiclass Classification model using any classification algorithm and create a model
- Read the test data and carry our data cleaning, encoding and vectorising operations on the test data
- Predict the sentiments for test data
- Print and explain the Confusion Matrix
- Compute Accuracy of your model