**DJ SYNAPSE: TASK 1(a)**

**NATURAL LANGUAGE PROCESSING**

**TITLE:** Sentiment Analysis on tweets fetched from dataset of tweets.

**WORK PLAN:**

1. Download the dataset from <https://www.kaggle.com/arkhoshghalb/twitter-sentiment-analysis-hatred-speech> .
2. Use data pre-processing to clean the dataset. This step is an integral part of an ML project.
3. The cleaned dataset contains words. Words cannot be used to train a machine learning model. Hence, we will use count vectorizer module of the famous scikit learn library of python to convert these words into a suitable numeric format.
4. After step 3 we will split the data into training and testing data.
5. We will use Support Vector classification algorithm to train the model on the training data.
6. Apply the model on the test data.

**PROJECT PLANNING:**

1. **Importing the Python Libraries:**

import numpy as np

import pandas as pd

import re

import nltk

nltk.download('stopwords')

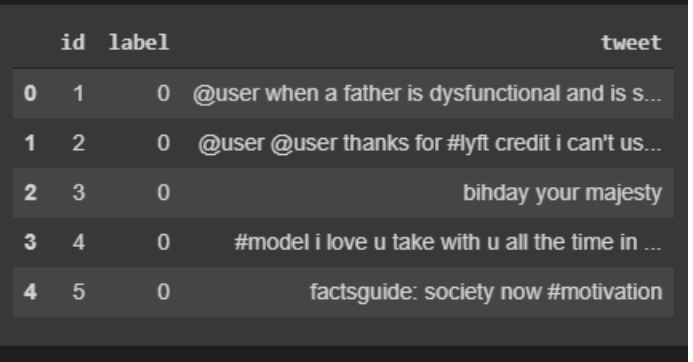
from nltk.corpus import stopwords

from nltk.stem.porter import PorterStemmer

import sklearn

1. **Dataset:**

The dataset contains two columns. The first column contains the label data, which is either “1” or “0”. “1” denotes the tweet in column 2 contains “racist” or “sexist” sentiment. The tweets column will be converted to predictor variable.



1. **Data Pre-processing:**

It is one of the most crucial steps for achieving greater accuracy for your model. As we can see from the screen shot the dataset, the tweets column contains textual data. We need to convert this data into a matrix of numbers. But before we do that we need to weed out the redundant word and symbols from the tweets and so that only the keywords that carry would express the desired sentiments remain. The following steps are taken :

1. **Convert all the letters to lowercase**.
2. **Tokenization:** Removing all the hashtags, numbers, @ or any other character other than the alphabets.
3. **Remove all the non-english words.**
4. **Removal of stop words:** Stop words are a collection of commonly used by a language that do not define the overall meaning of the sentence. But, depending on our goal, some stop word like “she”, ”not”, ”her” may play some part in understanding the sentiment as we are dealing with sexist tweets.
5. **Implementing the model:**

We will be using the Bag of Words model to train the data.

In this model, every new word encountered can be imagined to be stored in a Bag. A more fitting term maybe a “dictionary”.

Along with the words, it also stores the count of a particular word if it is encountered more than once. Hence, the removal of unwanted and redundant characters is important to improve the efficiency of the model.

The count vectorization module from the scikit learn library is used to create a number vector from the words. Each unique word holds an index position of the vector and the count of that word is stored as value of the vector corresponding to it’s index.

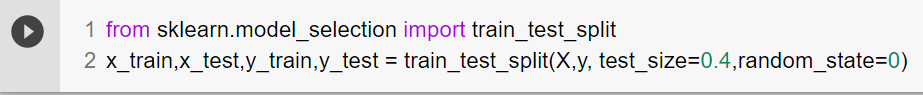
We store certain outlier words like proper nouns, at the last index.

A collection of such vectors from the whole dataset forms a 2D array.

1. **Splitting the Dataset into Training and Testing Data:**

We will assign 40% of the dataset to the training data

and the remaining 60% to testing data.



1. **Fitting the Support Vector classification algorithm:**

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N — the number of features) that distinctly classifies the data points.

We fit the training set on the classifier.

1. **Results and Accuracy:**

After the support vector classifier is trained we use it on the test set with the .predict() method. It returns an m-dimensional vector (m is the number of data points in training set) of 1’s or 0’s.

“1” represents sexist, racist sentiment in the tweet.

Since we, have the label data of the training set (y\_train) for reference, we can use a use the accuracy\_score() method from the sklearn.metrics module.

This method returns the fraction of correct predictions made by the model.