ASSIGNMENT DA-4

Roll No: 41205

Problem Statement:

Twitter Data Analysis: Use Twitter data for sentiment analysis. The dataset is 3MB in size and has 31,962 tweets. Identify the tweets which are hate tweets and which are not. Sample Test data set available here: https://datahack.analyticsvidhya.com/contest/practice-problem-twittersentiment-analysis/

Objective: Perform sentiment classification using Python with the Twitter dataset.

Outcome: One will be able to create a classifier to predict sentiments using necessary data.

Pre-requisites:

1. 64-bit Linux OS
2. Programming Languages: Python

Hardware Specification:

1. x86_64 bit
2. 2/4 GB DDR RAM
3. 80 - 500 GB SATA HD
4. 1GB NIDIA TITAN X Graphics Card

Software Specification:

1. Ubuntu 14.04

Theory:

The procedure to classify sentiments on the twitter dataset can be as follows:

- Remove stopwords:
    - Remove unnecessary words and symbols from tweets.
    - For eg. “This is a happy statement :)” would be converted to “This happy statement”.
- Stem words in the dataset:
    - To allow the model to understand between different forms of the same word.
    - For eg. work, worked, working. All of these words would be stemmed to the word “work”.
- Text vectorization:
    - Since the models can only understand numbers and words are strings, it is necessary to convert them to the numerical form.
    - This can be achieved by converting the dataset into a matrix of size *DICTIONARY SIZE x NUMBER OF TWEETS*.

    - Each sample can be converted into a one hot vector of size of the dictionary.
- Text classification:
    - The dataset can be passed as an input to a classification model with the labels.

Input: Twitter Dataset

Output:
Metrics

	precision	recall	f1-score	support
0	0.96	1.00	0.98	2984
1	0.88	0.46	0.60	213
accuracy			0.96	3197
macro avg	0.92	0.73	0.79	3197
weighted avg	0.96	0.96	0.95	3197

Confusion Matrix

```
[[2971   13]
 [ 116   97]]
```

Conclusion: Thus, we classified tweets into hate tweets and non-hate tweets using Python.