Assignment

- Title: Twitter Data Analytics.

- Problem Statement:

  Use twitter data for sentiment analysis. The dataset is 3MB in size and has 31962 tweets. Identify the tweets which hate tweets and which are not.

- Objective:

  - To apply sentiment analysis technique on twitter data and classify the tweets as hate or not.

- Outcome:

  Students will learn to apply proper prepocessing on twitter data and also learn the basics of sentiment analysis.

- Theory:

- A large amount of data that is generated today is unstructured, which requires processing to generate insights
  eg: data on news articles,
  posts on social media, etc.

- The process of analyzing natural language and making sense out of it falls under the field of Natural Language Processing (NLP)

- sentiment analysis is a common task which involve classifying texts or parts of texts into a predefined sentiment.

Dataset:

The dataset contains 31962 labelled tweets.

Different steps involved are:

(1.) Loading the data
(2.) Tokenizing the data
(3.) Normalizing the data
(4.) Determining word density
(5) Model training

1.) Loading the data:-

The data is provided as a csv file, we use 'pandas library' to load the data

2.) Tokenizing the data:

- Language in it's original form cannot be accurately processed by a machine to understand
- The first part of making sense of data is through a process called tokenization or splitting strings into smaller parts called tokens

3.) Normalizing data:

- Words have different forms - for instance, 'ran', 'runs' and 'running' are various forms of verb 'run'
- Normalizing in NLP is the process of converting word to it's canonical form
- Two popular techniques of normalization are stemming & lemmatization

4.) Determining word density:

- The most basic form of analysis on textual data is to take out the word frequency
- A single tweet is too small of an entity to find out the distribution of words, hence the analysis of the frequency of words would be done on all positive tweets

5.) Model training:

- We use the Naive Bayes Classifier to build the model.

- Analysis:

The train dataset was split to create a validation dataset.

The accuracy on the validation data is 83.25.% using Naive Bayes classifier.

- Conclusion:

We have successfully classified the twitter data into hate/not hate label with 83.25.% accuracy.