Date:-

Assignment No - C4.

Title:- Twitter Data Analysis.

Problem Statement:-

Use Twitter data for sentiment analysis the dataset is 3MB in size and has 31,962 tweets. Identify the tweets that are hate tweets and those that are not.

Objective:-

To classify tweets as hate tweets or not.

Outcome:-

Identifying and removing hate tweets from twitter.

Software and Hardware requirements:-

Python 3, Jupyter, pandas, numpy, sklearn, matplotlib; UNIX/LINUX based OS; 64 bit CPU with 8 GB RAM minimum.

Theory:-

Natural Language Processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence, concerned with interactions between computers and human

language, in particular how to program computers to process and analyze large amounts of natural language data.

Stop words are words that are filtered out before or after the natural language data are processed stemming for grammatical reasons, text can be use in different forms of a word. There are also families of derivationally related words with similar meanings.

Stemming reduces inflectional forms and sometimes derivationally linked forms of a word of its common base form.

When applied to a document, the result is like original, the boys ears are different colors STEMMED, the boy ear can be differ color.

Feature selection is the process of selecting a subset of the terms occuring in the training set and using only this subset of features in text classification

This makes the classifier more efficient, as well as more accurate because it eliminates noise.

Vectorization is the process are related to one-hot encoding, but instead of just featuring a count, they feature numerical representations where words aren't just present or not present - instead, they are represented by their term frequency multiplied by their inverse doc frequency.

For this particular problem, which is classifying tweets as hate tweets or not

The classification methods ~~used~~ were Multinomial Naive Bayes, Random Forest, and Linear Support Vector classifier.

Accuracy of >95% was achieved.

The tweets were pre-processed to convert them to lowercase, removed @ mentions, removed numbers, punctuation.

The tweets were the vectorized (TFIDF) and split into training & test data.

The 3 models were fitted and then used to predict the labels.