Rohit Kulkarni 41346

DA ASSIGNMENT 4

TITLE: Twitter Data Analysis

PROBLEM STATEMENT:
Use twitter data for sentiment analysis. The dataset is 3mb in size & has 31,962 tweets. Identify the tweets which hate tweets & which are not.

OBJECTIVES:
To classify tweets as hate tweets or not.

OUTCOMES:
To identify & remove hate tweets from twitter.

S/w & H/w REQUIREMENTS:
64 Bit OS, anaconda, jupyter notebook, nltk, pandas, matplotlib, packages, keyboard, mouse, monitor.

THEORY:

i] Natural language processing (NLP) is a subfied of linguistics, computer, science, artificial intelligence, concerned with interactions between computer & human language, in particular how to program computers to process & also analyse large amounts of natural language data.

ii] Stop words are the words that are filtered out before or after the natural language data are processed.

iii] Stemming for grammatical reasons, text can use different form of a word. There are also families of derivationally related words with similar meanings.

iv] When applied to a document, the result is like
ORIGINAL: the boy's cars are different colors.
STEMMED: the boy car be differe color.

v] Feature selection is the process of selection of a subset of the terms occurring in the training set & using only this subset of features in text classification.

vi] Vectorization is the process of converting the text data into a machine readable form.

vii] Accuracy of > 95% was achieved.

viii] The classification method used were Multinomial Naive Bayes, Random forest, Linear Support Vector Classifier.

ix] Accuracy ~~of~~ & tweets were prepossed to convert them to lowercase, removed the @mentions, removed numbers & punctuations.

CONCLUSION:

Tweets were successfully classified as hate tweets or not.