Assignment - DA4

* TITLE: Twitter Data Analysis.

* PROBLEM STATEMENT: Use Twitter Data for sentiment analysis. The dataset is 3 MB in size and has 31,962 tweets. Identify the tweets that are huge hate and tweets & those that are not.

* OBJECTIVE: To classify tweets as hate tweets or not.

* OUTCOME: Identifying & removing hate tweets from Twitter.

* SOFTWARE AND HARDWARE REQUIREMENTS: Python 3, Jupyter, pandas, numpy, S-klearn, matplotlib; UNIX/LINUX based OS; 64 bit CPU, 8GB RAM.

* THEORY:

• Natural Language Processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence, concerned with interaction between computers & human language, in particular how to program computers to process and analyze large amounts of natural lang data.

• 'stop-words are words that are filtered out before or after natural lang data are processed.

• Stemming: for grammatical errors reasons, text can use different forms of word. There are also families of derivationally related words with similar meanings.

• Stemming reduces inflectional forms and sometimes derivationally related forms of a word to its common base word.

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

- Feature selection is the process of selecting a subset of terms occuring in the training set & using only this subset of features in text classificati This makes classifier more efficient and accurate.

- Vectorization is the process of converting the text data into machine readable TF-IDF vectors 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.

- For this particular problem, which is classifying tweets as hate tweets.

- The classifications methods used were: Naive Bayes, Random forest, and Linear Support Vector classifier

- Accuracy of > 95% was achieved

- The tweets were preprocessed to convert them to lowercase, removed @ mentions, numbers & punctuations.

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

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

* CONCLUSION:

  Successfully classified tweets as hate tweets or not.