Roll no:- 41310

Page No.:

Date:

Assignment DA-4

* Title:- Twitter data analysis.

* 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.

* objective:-

Twitter data sentiment analysis for determing hate tweets & which are not.

* outCome:-

Students will be able to do sentiment analysis on twitter dataset using classification algorithm.

* Theory:-

Sentiment analysisis the process of determining wheather a piece of writing (Product/Movie review tweet etc) is positive, negative or neutral etc. It Can be used to identify the customer or followers attitude towards a brand through the use of variable such as context, tone, emotion etc. Marketers can use sentiment analysis to research public opinion of their Company & products.

* steps to perform sentiment analysis:-
* gather relevant tweets from twitter.
* preprocessing (stopword removal)
* feature extraction.
* feature selection.

1) Gather relevant tweets from twitter:-
various dataset are available of twitter data for sentiment analysis.

2) preprocessing:-
It makes raw text ready for mining and then it becomes easier to extract information and apply machine learning algorithm. If it is skipped we were working with noisy and inconsistent data. Now data will become less noisy and relevant for sentiment analysis. Punctuation, special characters, numbers & terms do not carry more weightage.

For preprocessing following steps are required:-
(A) Removing twitter handles (@user):-
The tweets contain lots of twitter handles (@user) that is how a twitter user acknowledged on twitter.

(B) Removing punctuations, Numbers and special characters punctuations, Numbers & special characters do not help much. It is better to remove them

(c) Removing short words:-

Some short words are removed for cleansing.

(D) Tokenization:-

Tokens are individual words and tokenization is the process of splitting a string of text into tokens.

(E) stemming:-

stemming is arule based process of stripping the suffixes ("ing", "ly". "es", "s" etc) from a word. for example- "Play", "Player", "Played", "Playing".

3) Feature Extraction:-

selection of useful words from the tweet is Called as feature extraction. In the feature extraction method extract the aspects from preprocessed twitter data set.

4) Feature Selection:-

Correct feature selection techniques are used in sentiment analysis that has important rules. They are Categorized into 4 main types, Natural language processing statistical clusturing based Hybrid.

5) ## Classification:-

(a) Naive Bayes Classifier

This approach uses Bayes theorem which describes how Conditional probability of each of set of possible causes for a given observed outcome

$$P(c|x) = \frac{P(x|c).P(c)}{P(x)}$$

← class prior probability.

↓ Posterior probability

→ Predictor prior probability.

(b) SVM Classifier:-

The support Vector Machine (SVM) is a algorithm to perform sentiment analysis. It analyzez data, define decision boundries and uses kernels for Computation.

Conclusion:-

Twitter sentiment analysis is performed to identify which are hate tweets and which are not.