Assignment DA-4

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 which are hate tweets and which are not

Objectives
Perform Twitter data sentiment analysis using Naive Bayes classifier

Outcome
To be able to do Sentiment analysis on Twitter dataset using classification algorithm

Software and hardware requirements
OS: 64 bit, open source Linux
Programming language : Python / R

Theory related concepts

Sentiment analysis is the process of determining whether a piece of writing (movie review / tweet) is positive, negative or neutral. It can be used to identify the customers or followers attitude towards a brand through the use of variables like context, tone, emotion etc. Marketers can use sentiment analysis to research public opinion of their

company and products or to analyze customer satisfaction. Organisations can also use this analysis to gather critical feedback about a problem in newly released products.

Steps to perform sentiment analysis

1. Gather relevant tweets from twitter
2. Preprocessing (Stopword removal)
3. Feature extraction
4. Feature selection

Preprocessing

The preprocessing of text data is an essential step as it makes the raw text ready for processing i.e. it becomes easier to extract information and apply machine learning to it. If we skip this step then there is a higher chance that you are working with noisy / inconsistent data. The objective of this step is to clean noise to find sentiment of tweets such as punctuation, special characters, numbers, terms which don't carry much weightage etc.

Initial data cleaning requirements

1. The twitter handles are already masked as @user due to privacy concerns.
2. We can also get rid of punctuations, numbers, special characters since they wouldn't help in differentiating different kinds of tweets.
3. Most of the small words do not add much value will also be removed.
4. Then we split, every tweet into tokens.

Stemming : Rule based process of stripping the suffixes from a word

Feature extraction

In this method, extract the aspects from preprocessed dataset.

1. There are different types of features namely unigram, bigram, n-gram features
2. Parts of speech tags such as adjectives, adverbs, verbs, nouns are good indicators of subjectivity and sentiment.
3. Negation is another important but difficult feature to interpret. The presence of a negation usually changes the polarity of the sentiment
4. Feature selection

Relevant attributes must be identified for increasing classification accuracy.

Techniques

1. NLP
2. Clustering
3. Statistical techniques
4. Hybrid techniques

Classification

Algorithm : Naive Bayes Algorithm

This algorithm uses simple approach based on Bayes Theorem which describes how the conditional probability of each of set of possible causes for a given observed outcome can be computed from knowledge of conditional probability of outcome of each cost./cause.

$$P(C|x) = \frac{P(x|C)\,P(C)}{P(x)}$$

$P(x|C)$ → likelihood, $P(C)$ → class prior probability, $P(C|x)$ → posterior probability, $P(x)$ → predictor prior probability

$$P(C|x) = P(x_1|C) \times P(x_2|C) \times \dots P(x_n|C) \times P(C)$$

The classification is conducted by deriving the maximum posterior which is the maximal $P(C_i|x)$ with above assumption applying to Bayes theorem. This assumption greatly reduces the computational costs by only counting the class distribution.

Test cases

	Input	Actual o/p	Expected o/p	Result
1.	Retweet if you agree	Positive	Positive	Pass
2.	It was a rough day at work today	Negative	Negative	Pass
3.	So excited for my birthday	Positive	Positive	Pass

Conclusion

Hence we successfully performed sentiment analysis on twitter dataset