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.

S/W and H/W Requirements:

Operating System : 64-bit Open source Linux or its derivative
Programming Languages: PYTHON/R

Learning Objective: Learn steps to do sentiment analysis

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

Concept related theory:

Sentiment Analysis is the process of determining whether a piece of writing (product/movie re view, tweet, etc.) is positive, negative or neutral. It can be used to identify the customer or fol lower's attitude towards a brand through the use of variables such as context, tone, emotion, etc. Marketers can use sentiment analysis to research public opinion of their company and products, or to analyze customer satisfaction. Organizations can also use this analysis to gather critical feedback about problems in newly released products.

Steps to perform sentiment analysis:

- Gather relevant tweets from Twitter
- Preprocessing (stopword removal)
- Feature Extraction
- Feature Selection

1. Preprocessing:

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

2. Initial data cleaning requirements:

- The Twitter handles are already masked as @user due to privacy concerns. So, these Twitter handles are hardly giving any infor information about the nature of the tweet.
- Most of the smaller words do not add much value. For example, 'pdx', 'his', 'all'. So, we will try to remove them as well from our data.
- In the 4th tweet, there is a word 'love'. We might also have terms like loves, loving, lovable, etc. in the rest of the data. These terms are often used in the same context. If we can reduce them to their root word, which is 'love', then we can reduce the total number of unique words in our data without losing a significant amount of information.

A) Removing Twitter Handles (@user)

As mentioned above, the tweets contain lots of twitter handles (@user), that is how a Twitter user acknowledged on Twitter.

B) Removing Punctuations, Numbers, and Special Characters

As discussed, punctuations, numbers and special characters do not help much. It is better to remove them from the text just as removed the twitter handles.

C) Removing Short Words

D) Tokenization

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

E) Stemming

Stemming is a rule-based process of stripping the suffixes ("ing", "ly", "es", "s" etc) from a word. For example, For example – "play", "player", "played", "plays" and "playing" are the different variations of the word – "play".

3. Feature Extraction:

Selection of useful words from the tweet is called as feature extraction. In the feature extraction method, extract the aspects from the pre-processed twitter dataset.

1. There are three different types of features namely unigram, bigram, n-gram features.

2. Parts Of Speech Tags such as like adjectives, adverbs, verbs and 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:

Correct feature selection techniques are used in sentiment analysis that has got a significant role for identifying relevant attributes and increasing classification (machine learning) accuracy. They are categorized into 4 main types namely,

- Natural language processing
- Statistical
- Clustering based
- Hybrid

A. Natural language processing mainly works on (1) Noun, noun phrases, adjectives, adverbs (2) Terms occurring near subjective expressions can act as features.

B. Clustering based feature extraction techniques are implemented by requiring few parameters. The major weakness of clustering is that only major features can be extracted and it is difficult to extract minor.

C. Statistical techniques are further divided into three sub types; they are univariate, multivariate and hybrid. Univariate methods, they are also called feature filtering methods, that take attributes separately, examples of this type include information gain (IG), chi–square, occurrence frequency, log likely-hood and minimum frequency thresholds. Univariate techniques have computational efficiency. Decision tree models, recursive feature elimination and genetic algorithms are the examples of multivariate methods. When compared to univariate; multivariate methods are expensive in terms of computational efficiency. Hybrid techniques are the one which combine the univariate and multivariate to achieve an efficient and accurate answer.

D. Hybrid techniques include POS Tagging with Word Net dictionary. Compactness and redundancy pruning methods were used for eliminating irrelevant features .

5. Classification:

Naive Bays Classifier: This classifier uses simple approach based on Bayes Theorem which describes - how the conditional probability of each of a set of possible causes for a given observed outcome can be computed from knowledge of the probability of each cause and the conditional probability of the outcome of each cause. It is a Bag of Words approach for subjective analysis of a content

. Naive Bayes (Classifier):

- Naive Bayes is a probabilistic classifier inspired by the Bayes theorem. Under a simple assumption which is the attributes are conditionally independent.

Test Cases:

Sr No.	Input	Actual output	Expected Output	Status
1.	is upset that he can't update his Facebook by texting it... and might cry as a result School today	Negative(0)	Negative(0)	Success
2.	Still procrastinating. .. i hate organizing my clothes there's just so much....	Negative(0)	Negative(0)	Success

Conclusion:

Thus we have successfully analyzed twitter data for negative and positive tweets and classified them appropriately too.