DSC 680 -PROJECT PROPOSAL MILESTONE 1

Abhijit Mandal

TOPIC – TWITTER SENTIMENT ANALYSIS

This project will focus on performing a sentiment analysis on a specific product or service which will be chosen as the project progresses. Opinions will be mined predominantly from twitter. The main goal of this sentiment analysis is to discover how users perceive the chosen product or service. The opinions that are mined will be classified into three categories positive, neutral and negative. An analysis will then be performed on the classified data to see what percentage of the population sample fall into each category.

BUSINESS PROBLEM

Sentiment analysis can be defined as a process that automates mining of attitudes, opinions, views and emotions from text, speech, tweets and database sources through Natural Language Processing (NLP). Sentiment analysis involves classifying opinions in text into categories like "positive" or "negative" or "neutral". It's also referred as subjectivity analysis, opinion mining, and appraisal extraction.

The words opinion, sentiment, view and belief are used interchangeably but there are differences between them.

- Opinion: A conclusion open to dispute (because different experts have different opinions)
- View: subjective opinion
- Belief: deliberate acceptance and intellectual assent

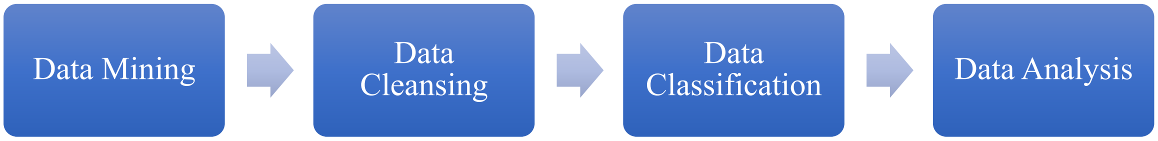
- Sentiment: opinion representing one’s feelings

Sentiment analysis and Natural Language processing are very important area nowadays. There is a massive amount of information being uploaded to the internet daily on social media websites and blogs that computers cannot understand. Traditionally it was not possible to process such large amounts of data, but with computer performance following the projections of Moore’s law and the introduction of distributed computing like Hadoop or Apache Spark, large data sets can now be processed with relative ease. With further research and investment into this area, computers will soon be able to gain an understanding from text which will greatly improve data analytics and search engines.

A customer’s perception or a product is extremely valuable data to some companies. From the knowledge gained from an analysis such as this a company can identify issues with their products, spot trends before their competitors, create improved communications with their target audience, and gain valuable insight into how effective their marketing campaigns were. Through this knowledge companies gain valuable feedback which allows them to further develop the next generation of their product.

METHODS

This section will highlight the technical approach that will be followed for this project and will include the system description.



Data Mining

The data (tweets) will be collected using twitters API. **Twitter API** is an open source library for python that allows access to the full range of twitters RESTful API functionality. It will be used to access Tweets and the results will be compared against the parsing method.

Data Cleansing

A tweet contains a lot of opinions about the data which are expressed in different ways by different users .The twitter dataset used in this project work is already labeled into two classes viz. negative and positive polarity and thus the sentiment analysis of the data becomes easy to observe the effect of various features. The raw data having polarity is highly susceptible to inconsistency and redundancy. Preprocessing of tweet include following points,

- Remove all URLs (e.g. www.xyz.com), hash tags (e.g. #topic), targets (@username)
- Correct the spellings; sequence of repeated characters is to be handled.
- Replace all the emoticons with their sentiment.
- Remove all punctuations, symbols, numbers.
- Remove Stop Words
- Expand Acronyms (we can use a acronym dictionary)
- Remove Non-English Tweets

The second phase of the system will be to cleanse the data collected, this will involve removing any punctuations and making everything lower case. This will help in the next stage of the project especially in the "Bag of Words" approach. Removing lower case words will decrease the redundancy in the database that will be used to store the words.

Classifying the data

Machine learning techniques require representing the key features of text or documents for processing. These key features are considered as feature vectors which are used for the classification task. Some examples feature that have been reported in literature are:

1. Words and Their Frequencies: Unigrams, bigrams and n-gram models with their frequency counts are considered as features.

2. Parts of Speech Tags: Parts of speech like adjectives, adverbs and some groups of verbs and nouns are good indicators of subjectivity and sentiment. We can generate syntactic dependency patterns by parsing or dependency trees.

3. Opinion Words and Phrases: Apart from specific words, some phrases and idioms which convey sentiments can be used as features. e.g. cost someone an arm and leg.

4. Position of Terms: The position of a term within a text can affect how much the term makes difference in overall sentiment of the text.

5. Negation: Negation is an important but difficult feature to interpret. The presence of a negation usually changes the polarity of the opinion.

6. Syntax: Syntactic patterns like collocations are used as features to learn subjectivity patterns by many of the researchers.

This is expected to be the most difficult part of the project; it will entail looking at individual words or groups of words in a tweet and attempting to assign a sentiment to them. This is no easy task as it is very difficult for a computer to “understand” slang words and sarcasm.

“Bag of Words” Model The bag of words approach will involve building databases of positive, negative and neutral words. Each tweet will be broken up into individual words and then compared to the words in the databases. When there is a match, a counter will be incremented or decremented by a fixed amount depending on a weighting assigned. When this process is complete the counter will be used to classify the sentiment for example if the words in the tweet are largely positive the counter should be high.

Data Analysis

When the data is classified, there will have to be some kind of analysis performed on it. This may include simple percentages of customer satisfaction, or a more complex analysis could be performed such as comparing the customer sentiment on two similar products with the aim of finding a correlation between good sentiment and high sales of those products.

ETHICAL CONSIDERATIONS

User information is kept anonymous for all Tweets. Also, no demographic information is included for any user or group of users having same sentiment, this will help in negating the bias for any specific group of users representing any section of society.

CHALLENGES

Following are some of the challenges faced in Sentiment Analysis:

1. **Identifying subjective parts of text**: Subjective parts represent sentiment-bearing content. The same word can be treated as subjective in one case, or an objective in some other. This makes it difficult to identify the subjective portions of text. For example:
    a. The language of the Mr. John was very crude.
    b. Crude oil is obtained by extraction from the seabeds. The word “crude‟ is used as an opinion in first example, while it is completely objective in the second example.
2. **Domain dependence**: The same sentence or phrase can have different meanings in different domains. For Example, the word “unpredictable‟ is positive in the domain of

movies, dramas etc. but if the same word is used in the context of a vehicle's steering, then it has a negative opinion.

3. **Sarcasm Detection:** Sarcastic sentences express negative opinion about a target using positive words in unique way. Example:

   "Nice perfume. You must shower in it."

   The sentence contains only positive words but actually it expresses a negative sentiment.

4. **Thwarted expressions**: There are some sentences in which only some part of text determines the overall polarity of the document.
   Example: "This Movie should be amazing. It sounds like a great plot, the popular actors , and the supporting cast is talented as well" In this case, a simple bag-of-words approaches will term it as positive sentiment, but the ultimate sentiment is negative.
5. **Explicit Negation of sentiment**: Sentiment can be negated in many ways as opposed to using simple no, not, never, etc. It is difficult to identify such negations. Example: "It avoids all suspense and predictability found in Hollywood movies." Here the words suspense and predictable bear a negative sentiment.

REFERENCES

1. Alec Go, Lei Huang, Richa Bhayani, 2009. Twitter Sentiment analysis, s.l.: The Stanford Natural Language Processing Group.
2. https://developer.twitter.com/en/docs/tutorials/how-to-analyze-the-sentiment-of-your-own-tweets
3. https://textblob.readthedocs.io/en/dev/quickstart.html#sentiment-analysis
4. https://towardsdatascience.com/sentiment-analysis-of-tweets-167d040f0583
5. Alexander Pak, Patrick Paroubek, 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining, s.l.: LREC.
6. Berry, N., 2010. DataGenetics. [Online]
   Available at: http://www.datagenetics.com/blog/october52012/index.html [Accessed 14 04 2014].