Sentiment Analysis on Twitter Data for product evaluation

[1]Prof. SudarshanSirsat, [2]Dr.Sujata Rao, [3]Dr.Bharti Wukkadada
[1]*(Assistant Professor IT, K. J. Somaiya Institute of Management Studies and Research)*
[2]*(Associate Professor IT, K. J. Somaiya Institute of Management Studies and Research)*
[3]*(Assistant Professor IT, K. J. Somaiya Institute of Management Studies and Research)*

Abstract: *as more and more people are expressing their views and opinions on various microblogging websites there has been a surge of data generated by the users, these websites have people sharing their thoughts daily because of short and simple form of expressions. We can consider such type of data as a resource and perform sentiment analysis on data of various products and services making better data driven decisions. This paper highlights the usefulness of sentiment analysis along with the type of data that is being analyzed, the complex process involved in analyzing the data, the different approaches that can be used, and an experimental observation using the Machine Learning approach*

Keywords: *Microblogging, Twitter, Sentiment Analysis, API, NLTK, Naïve Bayes Classifier Machine Learning Approach, Lexicon based Approach.*

I. Introduction

Sentiment Analysis with the help of text mining technique for identifying the sentiments of the humanfora product or service, also sometime referred as opinion mining involves creating a system to collect reviews about a product and categorize them as positive, negative and neutral. It helps market researchers evaluate the performance of the product, estimate the success of the product. All the categories of reviewsfor a product/service can be analyzed. For examplea review for a digital camera can be positive based on picture quality, but can be negative depending on how heavy it is.

This kind of analyzed information in systematic way, helps the market researchers with a clear picture of public sentiments for the product, similar to surveys as the data is created by customer.

Sentiment analysis is an algorithmic process where a word can be positive in one situation and can be negative in another situation. For example the word "long". If a customer said a mobile phone's battery life was long, that would be a positive opinion. If the customer specifies that the mobile phone's boot time was long, that would be is a negative sentiment. As a result we need to create a new system to analyze opinions for each type of product/service experiences. Also, people can show contradiction of opinions in their statements. Reviews can be interpreted on the basis of both positive and negative sentiments one at a time depends on how data is collected/gathered.

II. Data Description

Twitter is the most popular social networking platform over which people express their views and opinions about various trending topics with the help of short messages called as tweets. These tweets are text messages with maximum length of 140 characters, because of this short message service people make the use of acronyms, emoticons and other special characters with special meanings. A brief description of all the terminologies associated with tweets.

Emoticons: A replacement to a group of words with a pictorial representation that express human emotions like humor, temper, happiness, sadness etc.

Target: The tweet messages are sometimes targeted to a user letting them know about something or expressing an opinion about the target user. The "@" symbol is used within the tweet to address the target.

Hashtags: Hashtags within the tweets are used to make the tweet visible under a twitter topic, there can be multiple hashtags within a tweet, making a tweet visible under multiple topics.

If we consider the tweets a source of information, we need to materialize this information so that it can be extract some gains out of it.

III. Process Involved in Analysing sentiment data

The complex process of measuring the sentiments of the tweets involves 5 major steps:

1) Data Extraction: This step involved in sentiment analysis consists of gathering the data from social network site twitter source using tweepy API provided by python. The tweepy API not only helps in extracting the tweet text but also provides extra information about the tweets like likes and retweets. The data extracted from the

twitter topic is highly disorganized and contains different types of emojis, stop words and is not specific to a language hence automation involved in analysis of data is the best approach for text classification.
2) Text Cleaning: After the tweets for a topic are extracted before passing it to the classifier, we need to clean the dataset to remove emoji's, stop words so that the non-textual content not pertinent to the analysis is identified and removed.
3) Sentiment Analysis: Once the data is cleansed it's ready for classification, into positive, negative and neutral tweets. There are various approaches to sentiment analysis like Machine Learning, Lexicon based and Hybrid approach. Also there are some other approaches like Natural Language Processing and Nero Linguistic Programming. Machine Learning involves a training dataset and a testing dataset, where we used the training data and train the classifier using one of many algorithms like Bayesian Networks, Naïve Bayes classification, Maximum Entropy, Networks Support Vector Machine. The testing dataset is used to test the classifier for its accuracy in the tweets. Lexicon approach does not use any training dataset, it makes the use of an inbuilt dictionary where all words are associated with a human sentiment. The Hybrid combines the Machine Learning and the Lexicon approach to improve performance of the sentiment classifier.
4) Presentation Output: sentiment-analysis is a main aim to generate meaningful information of raw data. After the analysis is complete, we can perform visualizations like creating bar graphs, Time series and pie charts. Bar graphs can be used to measure the sentiment the tweet as positive, negative and neutral. Time Series can be used to measure the likes, retweets and average length of the tweet over a period. Pie charts can be used to analyze the source of the tweet.

IV. Approaches to sentiment classification

Sentiment Classification can be performed by classifying a feature in favorable as positive, as well as unfavorable as negative too. Will classified as 3 levels:
1) Document-Level Classification: A document level classification helps in classifying an entire opinion document into a positive or negative sentiment. When we are providing reviews for products or services the entire review helps in expressing positive or negative sentiments of the consumer towards the product.
2) Sentence-Level Classification: Sentence Level Classification involves breaking down a review into sentences, calculating the polarity for each sentence and then accordingly calculating the sentiments for each review.
3) Aspect-Level Classification: Aspect Level Classification judges various aspects of the entity and giving different opinions about different aspects, it does not focus on the language construction but focuses more on the opinion itself. The classification focuses around breaking an opinion into sentiment of an opinion and target of opinion.

V. Approached Used

- Twitter Application was created and twitter data for "Product Name" was crawled using tweepy API.
- Extracting information about the tweets like Likes and Retweets and storing information about the tweets in the data frame
- Tweet Data was cleaned to remove punctuations, digits, and whitespaces & converted them into vectors from string.
- TextBlob package performs sentiment analysis on the data by calculating the polarity of the tweet (from -1 to 1) using NLTK (Natural Language Processing toolkit) corpora and the movie review dataset trained using Naive Bayes Algorithm.
- We then analyze the tweets to categorize positive and negative words from the tweets and calculate a count of total positive and negative words and store the frequency of different positive and negative words.
- Next, we remove stop words and create a word cloud to analyze the sentiment.
- Finally, we calculate a sentiment ratio which is nothing but negative to positive word ratio, to determine the overall sentiment about the product is positive or negative.

VI. Sentiment Classifier

The Sentiment Classifier is created using the movie Review dataset where the reviews are marked positive and negative.
The feature extraction is done from positive and negative reviews and the data is trained using a Naive Bayes Classifier.

Naive Bayes Classifier
Naïve Bayes classification algorithm is a basic classification algorithm which assumes that classification of entities based on their attributes and attributes are independent of each other without any correlation between them.

It is 'Naïve' because the probabilistic calculation for each hypothesis are simplified by calculating the value of each attribute independently without considering the conditional dependencies between various attributes.

Let us consider Hypothesis (hypo) may be class that can be assigned to data instance (data).

One of the simplest ways to select high probability hypothesis is by making the using of prior knowledge of the problem. A way provided by the Bayes Theorem for probability calculation for a given hypothesis quantified by way of:

P (hypo | d) = (P(d | hypo) * P(hypo) / P(d))

Where

P (hypo | d): the hypothesis h as probability given data d is true → probability of posterior.

P (d | hypo): probability of data d given hypothesis hypo is true.

P (hypo): the probability of hypothesis being true (regardless of data)→ the prior probability of hypo.

P (d): Probability of data.

The posterior probability for several different hypothesis are calculated, and the highest probability formally called maximum a posteriori (MAP) hypothesis is selected. This can be written as:

MAP (hypo) = max ((P (d | hypo) * P (hypo)) / P (d))

VII. Experimental Observations

1) Twitter Topic : Cheerio's (A General Mills Product)
2) Number of tweets extracted : 200
3) Length of the tweets

Average length of tweets:70.105

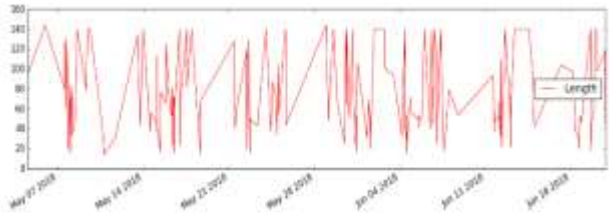

Figure 1. Length of tweets

4) Length v/s Likes v/s Retweets

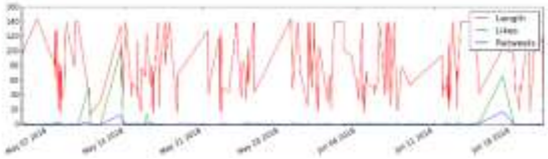

Figure 2. Length v/s Likes v/s Retweets

5) Positive v/s Negative v/s Neutral tweets

Percentage of positive tweets: 51.5%

Percentage of neutral tweets: 34.5%

Percentage of negative tweets: 14.0%

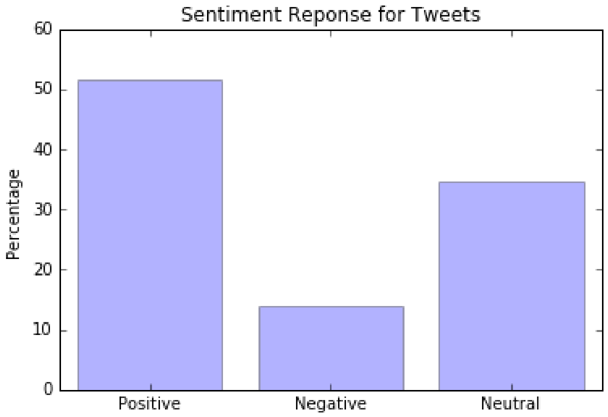

Figure 3. Positive v/s Negative v/s Neutral

VIII. Conclusion

The increase of various social platforms of twitter where people can use short messages to express their views and opinions helps us to create technologies which can help in analyzing for sentiments. This paper used the Naïve Bayes Algorithm to train the Movie Review Dataset, also uses using the TextBlob package in python to calculate the sentiments of the tweets. Along with the sentiments of the tweets we are also able to extract various characteristics of the tweets e.g. Likes, Retweets. The most liked tweet and the number of times it is retweeted. The accuracy in classification can be improved by using better models which can be trained using a larger dataset. The process thus defined is exploratory and we can improve it by using better approaches and algorithms.

References

[1]. Niu, Y., Zhu, X., Li, J., Hirst, G. 2005: Analysis of polarity information in medical text. In Proceedings of the American Medical Informatics Association 2005 Annual Symposium

[2]. AlessiaD'Andrea, Fernando Ferri, PatriziaGrifoni, TizianaGuzzo: Approaches, Tools and Applications for Sentiment Analysis Implementation

[3]. Balahur, A., Kozareva, Z., Montoyo, A. 2009: Determining the polarity and source of opinions expressed in political debates.

[4]. Medhat, W., Hassan, A., Korashy, H. 2014. Sentiment analysis algorithms and applications: A survey, Ain Shams Eng.

[5]. Apoorv Agarwal BoyiXie Ilia Vovsha Owen Rambow Rebecca Passonneau : Sentiment Analysis of Twitter Data

[6]. S. Karthika* and N. Sairam :A Naïve Bayesian Classifier for Educational Qualification.

[7]. Jebaseeli, A. N., &Kirubakaran, E. 2012. A survey on sentiment analysis of (product) reviews. International Journal of Computer Applications, 47(11).

[8]. Kaur, A., & Gupta, V. 2013. A survey on sentiment analysis and opinion mining techniques. Journal of Emerging Technologies in Web Intelligence, 5(4), 367-371.