ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 2, Issue 5, May 2013*

# Twitter based Sentiment Analysis for Subject Identification

Jisha S Manjaly[1]

Department of Computer Science, Rajagiri School of Engineering and Technology, Kochi, India [1]

**Abstract**: Nowadays the social media such as blogs, Twitter, Facebook etc. are widely used for participatory information sharing and collaboration. The analysis of the dynamic opinion of users in these social media forums will helps to answer many questions from various business and research fields across the world. Sentiment mining is used to identify and extract the sentiment and other emotional states in the online text. This paper aims to study the subject identification by analyzing tweets from Twitter, a micro-blogging site.

**Keywords**: Sentiment mining, Twitter, Naive Bayes, Fisher, API, Bag of Words, Bag of N-Grams**.**

## I. INTRODUCTION

Sentiment mining is a computational technique for extracting and assessing the opinions from various user generated text. It aims to determine the attitude of a speaker with respect to some topic. The social networking sites offer huge volumes of user generated data. These sites offer opinion of users from different political, geographical and social status. The dynamic nature of user generated data in the social networking sites makes them competent partner in the sentiment mining field.

The content and opinions expressed in the social media can helps to find the public opinion [3]. This extracted information might be the input for business and e-commerce applications. The trend of a product in a particular place or across the world can be determined from the opinion of the customer. Identification of the sentiments from the available data is the basic task associated with sentiment mining process. This may be the identification of the polarity of the words or sentence such as “positive”, “negative” and “neutral” or the emotional states such as “sad”, “happy” or “angry”.

Twitter is an online social networking and micro-blogging service. Twitter is sharing information between users through messages known as “tweets”. Each tweet can consist of maximum 140 characters. As of 2012, twitter has more than 500 million registered users and generating over 340 million messages per day [4]. So the data from the twitter helps us to judge the behaviour of the market.

A company can understand the trend of their product by analyzing the data captured from Twitter and that directs the market of a product. According to the customer satisfaction, the company can change the strategies about the product reliably. This paper focuses on the subject identification from the opinion of different users or customers with the help of data collected from the Twitter.

## II. BACKGROUND STUDY

Customer sentiment is the feelings that consumers have about a product. Sentiment is driven by the consumer experience as well, with people sharing their feelings about the use of a product or service. Sentiment mining is a subfield of natural language processing. It has to deal with the complexity of natural language. In many circumstances sentiment is expressed through simple constructions, e.g. "Honda Amaze is amazing”. But some simple constructions can cause difficulty depending on document representation. For example: "I totally refuse to accept that this phone is bad", contains the phrase "This phone is bad" yet the overall sentiment is exactly the opposite.

Document preparation [1] is one of the most important parts of Natural language Processing. It deals with the selection of what features to use to represent a document. A full text string representation is not very useful, because there is no easy method for finding the abstract similarity of two document-length strings.

One of the simplest and most common string representations is the Bag of Words model [1]. This model ignores the ordering of words. So the structure of the sentence is ignored and represents the document as counts of the number of occurrences of the words in the document. This leads to the loss of fine grained information in sentences, as in the case of “I don't like him, I like her." and “I don't like her, I like him”. Despite this obvious loss of information, the bag of words model is still common in a large array of applications, and performs very strongly. It is computationally simple and in many applications much of the information required for learning is captured by this representation.

A logical extension to the Bag of Words model is Bag of N-grams [1] [2]. In this representation, instead of storing counts of the occurrences of individual words, we store counts for groups of consecutive words of size n. This eliminates important ambiguities that are seen in bag of words models, such as “white house" being significantly different to “white horse was outside the house”. The advantage of increasing the length is that more contexts are captured by the representation.

Once the document preparation method is finalized, the next step is to convert the document to sentiment prediction. Majority of applications are used supervised learning techniques for this purpose. That is, we need to have a training set of documents D with their sentiment identified by a human expert. These examples are used to learn a model of how the documents map to the given sentiments and use this to predict the sentiment of new documents without labels. To do this, standard supervised learning techniques, such as Naive Bayes and Fisher Algorithm are used.

### A. *Naïve Bayes*

One of the simplest supervised learning techniques for predicting discrete outputs from discrete inputs is Naive Bayes. The defining feature of Naive Bayes [1] is it assumes that all features are independent. This is a very strong assumption, which is not true for most applications.

### B. *Fisher Algorithm*

Fisher Algorithm approach allows for a higher level of flexibility in identification of semantic combinations as opposed to the standard naïve Bayesian algorithm by supporting a normalized method of classification. To perform such normalization the method relies on the following calculations [1]:

$$\text{Clf} = \Pr(\text{feature}|\text{category})$$

$$\text{Freq sum} = \sum_{i\ldots}^{n} \Pr(\text{Feature}|\text{Category})$$

$$\text{Cprob} = \text{crlf}/(\text{clf}+\text{nclf})$$

Clf is determined as a conditional probability that a document fits into a category, given a particular feature. By approximating at the feature to category level, it takes into account of receiving far more documents in one category than another. To normalize, the probability is divided by the frequency sum. The Fisher method continues by multiplying all the probabilities together, taking the natural log and applying the inverse chi function to obtain a probability.

Next, the word support is calculated, identifying the important words within a category by using the following formula[1].

$$\text{Freq}(w,s)=N(w,s)/\sum_{s\epsilon \text{Negative,Neutral,Positive}} N(w,s)$$

Above, N(w,s) represents the number of words in a category thus determining the relative frequency of topics. The text samples are first filtered for a stop word removal. Next, the specific mutual analysis is applied to the word lists in each category. Following this step, the word distribution is calculated. Both scores are then added to form a composite score upon which the topic words are rank ordered.

## III. PROPOSED SYSTEM

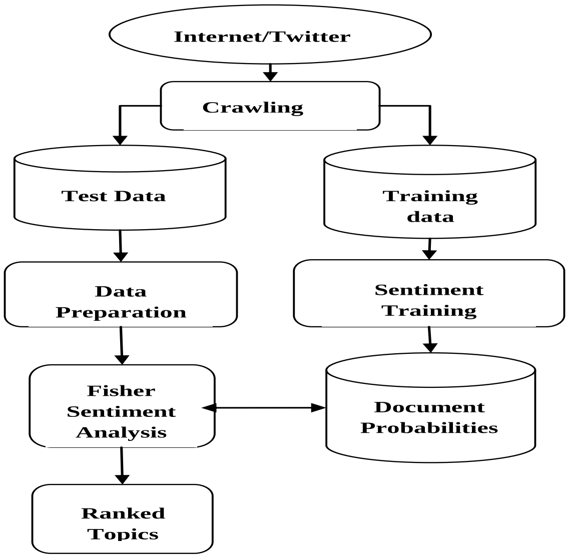

Fig 1. Sentiment Anlysis Model

The proposed system is shown in fig 1. Our process begins with an acquisition program applied to an internet-based data source (Twitter). Data is then retrieved from Twitter according to specified keywords. The data segments to be considered are then fragmented to support a derivation to a conceptual level. As applied to Twitter- based data, we make the assumption that each message will only contain a single concept. We have chosen three distinct categories towards our research goals. To each of the categories, we assign the following definitions:

NEGATIVE: negative sentiment directed towards product[1]
NEUTRAL: purely informational good or bad news (including headlines automatically forwarded through twitter accounts with no personal opinion)
POSITIVE: positive sentiment characterized as usually informational typified by positive informal (human generated) language.

### A. *Obtaining Data:*

Twitter provides two APIs to access information about tweets; the Search API[5] and the Streaming API. The Search API is intended to provide the ability to perform specific, low through-put queries. One can refine by user, content, geographic location and date. Importantly, only tweets from the preceding 5 days can be searched and queries are limited to approximately 10 per minute at the time of writing. The Streaming API is designed to allow access to a live stream of tweets, as they occur. One can refine by user, content and area. However, the user/content filter and the geographic filter are performed with a logical “OR”. This means that a search for tweets “Hyundai in Cochin” will return all tweets from Cochin and all tweets containing “Hyundai”. We are interested in predicting on the streaming API . Sentiment scores for each of the tweet using Fisher algorithm. The current model is capable of analyzing a maximum of 100 recent tweets related to the search criteria and displays the results to the user in CLI.

### B. DocumentPreparation

Once we have retrieved our tweets from the API, we convert them from the string representation to a feature representation. For our representation we choose a bag-of-words model. Tweets are not written like other text documents. The limitations of tweets such as maximum 140 character limit leads to lot symbols, short words, punctuations etc present in the messages compared to text documents. So a custom tokenizer and pre-processor are used for document preparation stage.

### C. Sentiment Analysis process

For the sentiment analysis, a training set of documents are manually devised according to our established classification definitions. Towards support of semantics identification we employ the Fisher classification method. This approach allows for a higher level of flexibility in identification of semantic combinations as opposed to the standard naïve Bayesian algorithm by supporting a normalized method of classification.

## IV. EXPERIMENTS AND RESULS

A sample model (Sentiment analysis tool) of the proposed solution is implemented using JAVA and MySQL[6] database. A third party package Twitter4J[5], a reference library for twitter in java, is used for crawling the data from the twitter. The Sentiment analysis tool first collects the search query from the user and connects the web for crawl the data from twitter. The crawled data is stored in to a MySQL database and find the sentiment score for each of the tweet using Fisher algorithm. The current model is capable of analyzing a maximum of 100 recent tweets related to the search criteria and displays the results to the user in CLI.

### A. Database

TABLE 1
DESCRIPTION OF TWEETS TABLE

| Field | Description |
|---|---|
| ID | Unique Id of the tweet |
| TEXT | The body text of the tweet |
| TIME | Time of the tweet |
| USER | The author of the tweet |
| SENTIMENT | Sentiment Score of the tweet data |
| DESCRIPTION | Description of the tweet |

The database used for the sample model is MySQL. Description of the sample table is shown in table 1. MySQL is the most popular open source RDBMS that runs as a server providing multi-user access to a number of databases. A database named twitterDB and a table named tweets has been created for storing the tweets crawled from the twitter.

### B. Sentiment Analysis Tool

The Sentiment Analysis tool consists of two jar files and a shell script; SearchTweets.jar, Generate SentimentInfo.jar and TwitterSentimentAnalysis.sh file. The sentiment analysis tool is implemented using JAVA. The tool is mainly composed of two sections; Data capture and Sentiment analyzer. The Data Capture section captures the input from the user i.e.; the search query and number of tweets to analyze, using the CLI and forward this data to the crawling unit. With the help of Twitter4J API, the crawling unit capture the recent tweets related to the search query from the web. The crawled data will be stored in to tweets table with the help of JDBC API.The Sentiment analyzer section collects the data from the tweets table and finds the sentiment score for each of the tweet using Fisher Algorithm. The results will be displayed to the user using the CLI.

#### 1) Running the Sentiment Analysis Tool:

Open the terminal and change directory to the location where the TwitterSentimentAnalysis.sh file is located and execute ./TwitterSentimentAnalysis.sh

#### 2) Output:

The sample input is the latest version of Samsung Galaxy series phone "Samsung Galaxy S4". The sentiment analysis tool search the query over internet within Twitter and find out the latest 100 tweets from different users across the globe and download the tweets in to the database. The sentiment info calculator calculates the sentiments associated with the words in each sentence and find out the total sentiment score of the sentence. Finally the tool will group the results in to 3 categories named negative, neutral and positive. The final results will be displayed to the user through Command Line Interface (CLI).

```
        ./TwitterSentimentAnalysis.sh
--------------Welcome to Twitter Search for Sentiment
Analysis*********Implemented By JISHA S
MANJALY*********--------------------------------------Enter
the search query : (Example Sachin Tendulkar)
Samsung Galaxy S4
Enter Maximum number of tweets to analyze : (1-100)100
Please wait...
Connecting to Twitter to crawl the data...
Crawling is over. tweets. Size() : 100
Clearing the table tweets for storing new dataTable Tweets
cleared
Adding tweet data to tweets table
100 rows added to tweets
tableFile Generation From database for sentiment
calculation Completed
Twitter Search completed.
Please wait...
Calculating the sentiment score for the tweets...
```

```
File generated with sentiment scoregetQueryString from file
is Done.
Sentiment calculation from generated files is ongoing...
Total Tweets Analysed => 100
Negative tweets => 2 ( 2.0 %)
Positive tweets => 34 ( 34.0 %)
Neutral tweets => 64 ( 64.0 %)
Sentiment analysis using Twitter data is completed.
```

## V. CONCLUSION

With the rapid growth of Internet and the evolution of web 2.0 websites, tradition marketing methodologies become harder and harder to satisfy the customers. Social networking and collaboration websites now are very popular internet applications and therefore attract more and more users. Thus, the social networking websites have also become a very good resource and platform for marketing. A combination of social network analysis and sentiment mining helps to identify the sentiments associated with the user opinion in social media. This paper proposes a method for integrating both techniques based on twitter. A system prototype has also been implemented in this paper with the help of Twitter to show how the mechanism works. In the future, we will apply the methodology to different data and resources in the internet. We will also focus on how to measure the system performance.

## ACKNOWLEDGMENT

The heading of the Acknowledgment section and the References section must not be numbered.

Causal Productions wishes to acknowledge Michael Shell and other contributors for developing and maintaining the IEEE LaTeX style files which have been used in the preparation of this template

## REFERENCES

[1] David Alfred Ostrowski System Analytics Research and Innovation Center Ford Motor Company, *Sentiment Mining within Social Media for Topic Identification*, 2010 IEEE International Conference on Semantic Computing

[2] D. Bespalov, B. Bai, Y. Qi, A. Shokoufandeh , *Sentiment Classification Based on Supervised Latent n-gram Analysis* ,the 20th ACM Conference information and Knowledge Management 2011.

[3] Lau, Raymond Y.K.; Lai, C.L.; Li, Yuefeng;, *Leveraging the web context for context-sensitive opinion mining*, 2nd IEEE International Conference on Computer Science and Information Technology, 2009. ICCSIT 2009.8-11 Aug. 2009 pp. 467 – 471

[4] Twitter. http://en.wikipedia.org/wiki/Twitter

[5] Twitter4J – A Reference Library for Twitter in Java. http://twitter4j.org/en/index.html

[6] MySQL –The world's most popular open source database. http://www.mysql.com/

[7] J. Padhye, V. Firoiu, and D. Towsley, "A stochastic model of TCP Reno congestion avoidance and control," Univ. of Massachusetts, Amherst, MA, CMPSCI Tech. Rep. 99-02, 1999.

[8] *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification*, IEEE Std. 802.11, 1997.