Sentiment Analysis for Social Media

Asst Prof. Indumathi S K[1] Chethan V[2]

1Professor of Dr Ambedkar *Institute of Technology, Dept of MCA, Bangalore-560056, Karnataka, India*
2Student of Dr Ambedkar *Institute of Technology, Dept of MCA, Bangalore-560056, Karnataka, India*

Abstract - Sentiment analysis, the automated extraction of expressions of positive or negative attitudes from text has received considerable attention from researchers during the past decade. In addition, the popularity of internet users has been growing fast parallel to emerging technologies; that actively use online review sites, social networks and personal blogs to express their opinions. They harbor positive and negative attitudes about people, organizations, places, events, and ideas. The tools provided by natural language processing and machine learning along with other approaches to work with large volumes of text, makes it possible to begin extracting sentiments from social media. In this paper we discuss some of the challenges in sentiment extraction, some of the approaches that have been taken to address these challenges and our approach that analyses sentiments from Twitter social media which gives the output beyond just the polarity but use those polarities in product profiling, trend analysis and forecasting. Promising results has shown that the approach can be further developed to cater business environment needs through sentiment analysis in social media.

Key words – Sentiment Analysis, Natural Language Processing, Data Mining, Supervised Learning

I INTRODUCTION

People make judgments about the world around them when they are living in the society. They make positive and negative attitudes about people, products, places and events. These types of attitudes can be considered as sentiments. Sentiment analysis is the study of automated techniques for extracting sentiments from written languages. Growth of social media has resulted in an explosion of publicly available, user generated text on the World Wide Web. These data and information can potentially be utilized to provide real-time insights into the sentiments of people [1].

Blogs, online forums, comment sections on media sites and social networking sites such as Facebook and twitter all can be considered as social media. These social media can capture millions of peoples' views or word of mouth. Communication and the availability of these real time opinions from people around the world make a revolution in computational linguistics and social network analysis. Social media is becoming an increasingly more important source of information for an enterprise. On the other hand people are more willing and happy to share the facts about their lives, knowledge, experiences and thoughts with the entire world through social media more than ever before. They actively participate in events by expressing their opinions and stating their comments that take place in society. This way of sharing their knowledge and emotions with society and social media drives the businesses to collect more information about their companies, products and to know how reputed they are among the people and thereby take decisions to go on with their businesses effectively. Therefore it is clear that sentiment analysis is a key component of leading innovative Customer Experience Management and Customer Relationship Marketing focused enterprises. Moreover for businesses looking to market their products, identify new opportunities and manage their reputation. As businesses look to automate the process of filtering out the noise, understanding the conversations, identifying the relevant content and take appropriate action upon it. Many are now looking to the field of sentiment analysis. In the era which we live today, sometimes known as information age, knowledge society; having access to large quantities of information is no longer an issue looking at the tons of new information produced everyday on the web. In this era, information has become the main trading object for many enterprises. If we can create and employ mechanisms to search and retrieve relevant data and information and mine them to transfer it to knowledge with accuracy and timeliness, that is where we get the exact usage of this large volume of information available to us.

However, in many cases these relevant data and information are not found in structured sources such as tables or databases but in unstructured documents written in human language. Human languages are ambiguous and the same sentiment can be used to express two different ideas in two different contexts. Moreover some people use different jargon, slang communications and short forms of the words for their ease. Therefore, it is difficult to gauge and measure the sentiments accurately in terms of their polarity such as positive, negative or neutral and the subjectivity of sentiments [2].

Most solutions in the market today rely on simple Boolean terms to express sentiment about a post, tweet, Facebook wall post etc. But this is not enough to address the above mentioned problems in the area of sentiment analysis and it will not generate precise and timely knowledge for aggregate sentiments. In order to get accurate knowledge after analyzing a sentiment, it should thoroughly consider solving the issues mentioned above. Most other systems that try to give solutions for these issues are still on research level, some systems also try to analyze sentiments from multiple languages and few systems which address some of the above mentioned drawbacks are available commercially also.

This paper reveals an approach which is implemented as a tool that can analyze sentiments on twitter social media addressing above issues and then develop an application to generate knowledge that can be useful for business environments using people's attitudes about their products and services.

II LITERATURE REVIEW

This section illustrates other similar work related to analyzing sentiments. Most of these approaches analyze sentiments as positive and negative while some approaches are in research level and few more are commercially available.

Adobe Social Analytics

Adobe Social Analytics basically measures the impact of social media on businesses by understanding how conversations on social networks and online communities influence marketing performance. After capturing and understanding the conversations going on, it correlates the impact of those conversations with key business matrices such as revenue and brand value. Other than that it measures the interactions that businesses have with their customers in social media including how Facebook posts drive site visitors and purchase behaviors [3]. Adobe Social Analytics uses a natural language processing algorithm to implement sentiment analysis.

Brandwatch Sentiment Analysis

Brandwatch is also a sentiment analysis tool developed by a team of PhD qualifiers in the United Kingdom; this is also commercially available currently. Through this tool they are trying to access whether a sentiment is positive, negative or neutral [4].

Sentiment140

This is an online tool for analyzing sentiments of Twitter social network. This tool allows discovering the sentiment of a brand, product or topic on Twitter. This was created by three Computer Science graduate students at Stanford University and their main focus is analyzing the languages English and Spanish. Sentiment140 basically states whether the specified brand, product or topic is positive, negative or neutral [5].

Social Mention

Social Mention is a social media search and analysis platform which analyses user sentiments through social media. This is also an online tool that allows tracking what people are saying about a particular brand, product or topic in real time. This tool allows the user to define a time period in which to analyze user sentiments.

TweetFeel

TweetFeel is also a web tool that analyzes sentiments of the given input through the twitter social media. This gathers real time data on Twitter, about the search items and evaluates those tweets into positive and negative categories in real time. This uses machine learning based sentiment analysis which enables to get much clearer feeling about sentiments.

Determining the Semantic Orientation of Terms through Gloss Classification

Sentiment classification is a recent sub discipline of text classification which is concerned not with the topic a document is about, but with the opinion it expresses. In this approach of sentiment classification it uses a method that is based on the quantitative analysis of the glosses of such terms, i.e. the definitions that these terms are given in on-line dictionaries, and on the use of the resulting term representations for semi-supervised term classification [6].

Sentiment Analysis using Adjectives and Adverbs

While most work in sentiment analysis determine its polarity using specific parts of speech such as adjectives, verbs and nouns, in this approach it uses Adverb-Adjective Combinations (AACs) to determine the strength of subjective expressions of a sentence. Instead of aggregating scores of both adverbs and adjectives using simple scoring functions, it proposes an axiomatic treatment of AACs based on the linguistic classification of adverbs. Three specific AAC scoring methods that satisfy the axioms are presented [7].

The specialty in our system is, it does not only analyze the sentiments, and instead it uses the analyzed sentiment scores to provide product profile, trend analysis and forecasting for the user.

III APPROACH

To analyze sentiments and then come to a conclusion through them, we need to have enough sentiments in the correct format. There are thousands and millions of sentiment data in the web, especially in social media sites that can be used to get valuable conclusions. But they are not in a correct format or not in a structured way to get maximum usage out of them. We need to convert them to a correct format and use them as we want. It is the first part in our approach, which is developing a crawler to crawl data from Twitter social media. The Crawler should be able to crawl user sentiments from twitter and at the same time get user details in order to do product profiling for customers as the later part of the whole approach.

After having access to large sources of data which is in a structured manner through the crawler and using a database, the next step is analyzing sentiments. Sentiments can be in different languages; in our project we cover the English language. Analyzing sentiments is a way of processing natural languages, therefore this part is about natural language processing. For this we use Natural Language Toolkit, also known as NLTK which is a leading platform for building Python programs to work with human language data. There are different ways that we can use to analyze sentiment data using this toolkit, but none of them gives hundred percent accuracy, because natural languages are used in many different ways by people. In our approach that is presented through this paper, we have implemented two supervised learning techniques named Naïve Bayes Classification and Maximum Entropy Classification to classify unknown sentiments, which later gives the probability of how much if a sentiment is positive and negative. As another method it uses SentiWordNet which is a lexical database that assigns scores to words and thereby finds a sentiment score to an entire sentence. It chose the first method which is Naïve Bayes Classification as the best method out of these three techniques after evaluating all the three methods. Using the selected sentiment analyzing method in this approach, it gives not only positive, negative and neutral sentiment score to the user sentiments, but it can solve the issues of using short words, different jargon words and smileys in social media. To differentiate the senses of ambiguous words such as “apple”, it used word sense disambiguation technique and could be able to differentiate different senses for ambiguous words.

Then as the third part of the project, we created a dashboard to show the results from the crawler and sentiment analysis using python. Here it will display how the sentiment polarity differs for a selected item with the time using a graph. Using this we can visualize how it is changing the user sentiment polarity of a specified brand or product with time. As the final part, the output of the sentiment analysis module will be used as the input for data mining module. It used the sentiment scores of a particular product or service with the user information such as age, profession, area and gender to profile products, analyze the trend for that particular product or service and forecasting. The result from data mining can be applicable in a most profitable way such as analyzing how people sentiments changes and will change for products and services with their age, location, profession and gender.

Anyone who is interested in searching what people say about a particular product or service can use this system. This is especially very useful for organizations producing products and services, to know what people are talking about their products and services and were they positive or negative, who are the competitors they have and what they can do to improve the reputation of their products and services among customers and what features should they include in order to attract positive remarks about their products.

IV DESIGN AND IMPLEMENTATION

There are four main modules naming, crawler, sentiment analysis tool, data mining module and dashboard in this project. The top level architectural diagram of the system as follows.

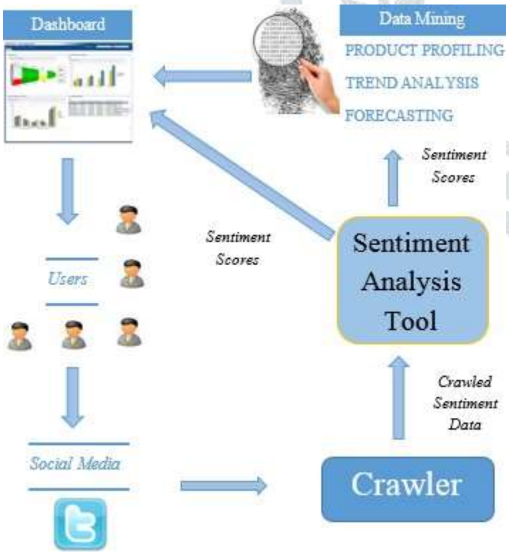


Fig 1: Top level architectural diagram of the system

The design and implementation details of those modules are described as follows.

Crawler

The basic purpose of the crawler is to gather social media data to a local data source for ease of analysis. For this it had to use a twitter API to get access to twitter data. There are a set of streaming APIs offered by Twitter which gives developers access to Twitters’ global stream of tweet data. Twitter offers several endpoints that have been customized to certain use cases such as Public streams, User Streams and Site Streams. Out of these three, for the crawler in this project we use the Public stream which is suitable for users or topics and data mining which reads the stream and directs those data to a database.

Sentiment Analysis Tool

This is the tool that analyzes user sentiments and gets the correct polarity of the given sentiments. Here it gets the data from the database that has been crawled; as inputs to this tool. Inside this tool, it uses different machine learning techniques to get the most accurate answers for the given sentiments thorough a proper classifier. The purpose of this classification task is to classify sentiments automatically basically into positive, negative and neutral categories which mean choosing the correct class label for a given input. Since it uses supervised classification, it is a prerequisite that there need to be a labeled text corpus into categories to train, test and build the classifier.

The next most important task after having a labeled text corpus is to find a way to extract features out of labeled corpus to train the classifier. The entire system depends on how good this method of feature extraction is. Therefore, it uses many different feature extraction methods in sentiment analysis as follows.

- Unigrams - Take each word individually in sentences as the feature set of corresponding category. Here it does not consider any relationship between words.
- Unigrams except Stop Words – Same feature extraction as above except it does not consider stop words which is a list of words that frequently appears in almost all sentences with no meaning.
- Bigrams – Take each adjacent two words in sentences as the feature set of corresponding category.
- Bigrams except Stop Words – This is same as Bigrams feature set, except words in Stop Words list.
- Most Informative Unigrams and Bigrams – Get the feature set with unigrams and bigrams with highest informative and highest frequency.

Out of these feature extraction methods, the last one which is most informative unigrams and bigrams were selected to use in our approach. Sometimes extracting too many features reduces the accuracy; therefore, in the above selected method it does not remove stop words because when it does, it reduces the accuracy. Then, extracted features need to be trained using a supervised classifier. As supervised classifiers, it uses Naïve Bayes classifier and Maximum Entropy classifier basically [8].

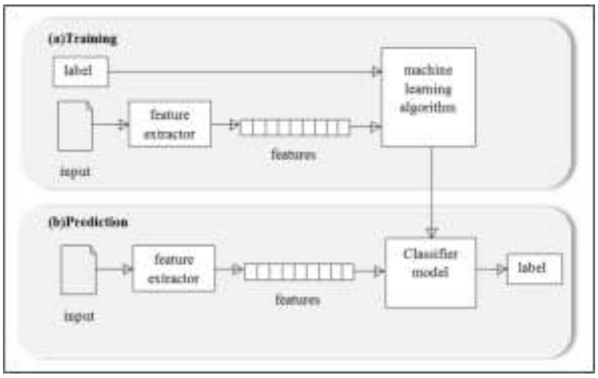


Fig 2: Supervised Classification

The above figure shows how generally a classifier is being trained and then predicts labels for unknown inputs. Identifying a better feature extraction method is the most important thing in building a classifier.

1. *Naïve Bayes Classifier*

The Bayesian classification is used as a probabilistic learning method. In naive Bayes classifiers, every feature gets a say in determining which label should be assigned to a given input value. To choose a label for an input value, the naive Bayes classifier begins by calculating the prior probability of each label, which is determined by checking frequency of each label in the training set. The contribution from each feature is then combined with this prior probability, to arrive at a likelihood estimate for each label. The label whose likelihood estimate is the highest is then assigned to the input value.

Following equations are used in calculating label likelihoods as shown in the following figure, Fig 2,

P(label|features) = P(features, label) / P(features)

= P(label) × P(features|label)) / P(features)

= P(label) × Prodf in| featuresP(f|label)` / P(features)
(Since features are independent)

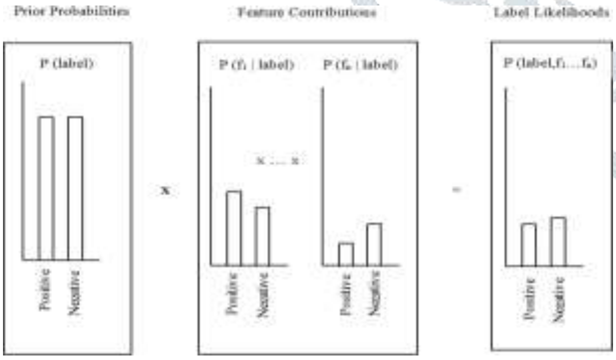


Fig 3: Calculating label likelihoods with Naive Bayes

In this classification method, the naïve Bayes assumption or independence assumption which is independence of the features makes it easy to combine the contributions of different features since it need not worry about how they should interact with one another.

2. *Maximum Entropy Classifier*

This classifier uses a model similar to naïve Bayes classifier, except it uses search techniques to find set of parameters to maximize the performance of the classifier rather than using probabilities to set model's parameters. Because of the potentially complex interactions between the effects of related features, there is no way to directly calculate the model parameters that maximize the likelihood of the training set. Therefore, Maximum Entropy classifiers choose the model parameters using iterative optimization techniques, which initialize the model's parameters to random values, and then repeatedly refine those parameters to bring them closer to the optimal solution [9].

3. *A Technique using SentiWordNet*

To implement this approach, it used set of labeled text corpus and SentiWordNet lexical data source which contains negative and positive scores for each word it contains. The following algorithm explains how we can analyze sentiments using the labeled corpus and the SentiWordNet. Here first it was needed to put the words of labeled corpus into two different bags called positive bag of words and negative bag of words based on the frequencies of each word.

```
pos = neg = 0
FOR sentence IN sentence_list :
   sent_score_pos = sent_score_neg = 0
   FOR word IN sentence:
      best_sense = Disambiguate Word Senses (sentence,
      word)
      IF bag_of_words['neg'].has_key(word):
         sent_score_neg += SentiWordNet[best_sense]['neg']
      IF bag_of_words['pos'].has_key(word):
          sent_score_pos += SentiWordNet [best_sense]['pos']
      pos += sent_score_pos
      neg += sent_score_neg
 RETURN pos, neg
```

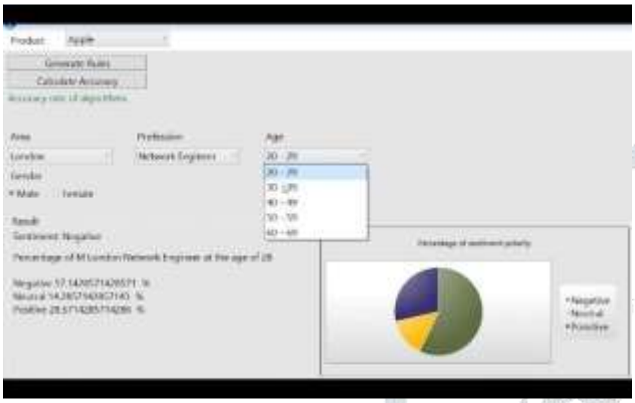
After training and testing these three methods with a labeled twitter corpus, we selected the Naïve Bayes classifier for sentiment analysis since it's faster and at the same time it gives little more accuracy than Maximum Entropy classifier and much more accuracy than the technique using SentiWordNet. The evaluation results of the two classifiers are stated in the evaluation section at the end.

Finally, as the solution to the problem of identifying ambiguous words in different contexts, we implemented word sense disambiguation to get the correct user sentiments with the intended sense. For this also we used Naïve Bayes classification method to classify a given set of ambiguous sentiments. For example, it requires disambiguating the sentiments of the product "Apple" from the fruit "Apple". Here, to correctly identify these two, it needs to do word sense disambiguation.

After implementing the sentiment analysis tool as mentioned above, we were able to overcome most problems that were encountered at the beginning of this approach. Then we were able to start the final part of the project that is implementing data mining module which uses the analyzed sentiment scores to product profiling and forecasting.

Data Mining

When implementing the product profiling, it uses the decision tree after comparing it with the clustering technique and for the trend analysis and forecasting it used Holt Winters method which is capable of analyzing seasonal data and predict proper values for the future. The following figures show how these product profiling, trend analysis and forecasting works in this system.


Product: Apple

Area: London

Profession: Network Engineer

Age: 20 - 29

Gender: ● Male ○ Female

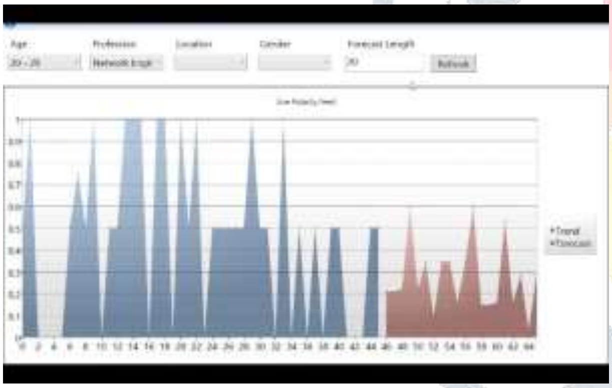
Result

Sentiment: Negative

Percentage of M London Network Engineer at the age of 28

Negative 57.1428571428571 %
Neutral 14.2857142857143 %
Positive 28.5714285714286 %

Fig 4: Product profiling


Age: 20 - 29

Profession: Network Engi

Location:

Gender:

Forecast Length: 20

Fig 5: Trend analysis and forecasting

V EVALUATION

Here we give priority to evaluate the classifiers in order to select the most probable classifier to use in the project Sentiment Analysis for Social Media. This evaluation was an effective tool to guide us to make improvements to the selected model also.

To evaluate these two classifiers, we needed a large labeled text corpus. Since it could find a very large corpus, it trained the classifiers on large set of training data and tested on increasing amount of test data. Prior to state the evaluation results of the classifiers, it needs to present evaluation results of feature extraction methods because it is the major thing that affects to the accuracy of the classifiers. The evaluation results of the feature extraction methods as well as the classifiers with selected feature extraction method are as follows.

TABLE 1
EVALUATION OF DIFFERENT FEATURE EXTRACION METHODS

Feature Extraction Method	Accuracy	Positive Precision	Positive Recall	Negative Precision	Negative Recall
Unigrams	0.7483	0.8058	0.6543	0.7090	0.8424
Unigrams except Stop Words	0.7436	0.7605	0.7112	0.7288	0.7761
Bigram Collocation	0.7630	0.8143	0.6814	0.7261	0.8447
Most Informative Unigrams	0.7778	0. 8581	0.6658	0.7269	0.8899
Most Informative Unigrams and Bigram Collocations.	0.7785	0.8350	0.6881	0.7347	0.8641

TABLE 2
DIFFERENT EVALUATION METHODS

Evaluation Method	Classifier	Training Set Size	Test Set Size
1	Naïve Bayes	70000	10000
2	Naïve Bayes	70000	20000
3	Naïve Bayes	70000	30000
4	Maximum Entropy	70000	10000
5	Maximum Entropy	70000	20000
6	Maximum Entropy	70000	30000

TABLE 3
EVALUATION RESULTS FOR ABOVE EVALUATION METHODS

Evaluation Method	Accuracy	Positive Precision	Positive Recall	Negative Precision	Negative Recall
1	0.7751	0.8357	0.6848	0.7330	0.8654
2	0.7758	0.8336	0.6891	0.7350	0.8625
3	0.7806	0.8388	0.6946	0.7394	0.8666
4	0.7730	0.8373	0.6780	0.7294	0.8680
5	0.7765	0.8388	0.6846	0.7335	0.8685
6	0.7799	0.8422	0.6889	0.7368	0.8710

After evaluating the results of the evaluation process, as the most suitable supervised learning method from accuracy wise and performance wise, it selected the Naïve Bayes classifier to classify sentiments with most informative unigrams and bigrams as the feature extraction method.

VI CONCLUSION

It is a very important fact to analyze how people think in different context about different things. This becomes more important when it comes to the business world because business is dependent on their customers and they always try to make products or services in order to fulfill customer requirements. So knowing what they want, what they think and talk about existing products, services and brands is more useful for businesses to make decisions such as identifying competitors and analyzing trends. Both because people express their ideas on social media and it can access those data, it has been enabled in some way to do the above mentioned things by using those data. The project, Sentiment Analysis for Social Media does that.

From the view at the top level of the project, we get data from social media sites to extract sentiments out of them and keep record of those sentiments with the information of the users who stated those sentiments in order to be used later. Finally it does data mining with the extracted sentiments so that it can be used in product profiling, trend analysis and forecasting.

After developing the crawler, the main challenge that was to be addressed was, how to decide whether a given sentence was positive or negative or neutral. The first thing that was found to address this challenge was a lexical data source which is called SentiWordNet, in that it has positive and negative score for each word. Though there are positive and negative scores for almost all words in English language, when it comes to sentences, it differs the overall polarity of a sentence with other words and according to the context. Other than that, it cannot analyze words with short terms which in returns reduce the accuracy and sometimes it makes the result incorrect. Moreover, it sometimes did not give correct polarity values for sentences which includes terms like ‘not good’, ‘not bad’.

During the implementation of the sentiment module we had to consider several issues such as, the comment by the user of a product or a brand can be not only in English but also mix with other language (Sinhala/Tamil), with emotional symbols etc., the comment may not completely match with what exactly user need to express about the product or brand, identifying the entity, identifying the relation of a particular comment with previous comments, ambiguity of words of the comment, human language is noisy and chaotic and the users may use different jargon or slang communications. But with the implementation machine learning techniques, it could achieve more accurate results after building classifiers training on large labeled data sets but still there are some issues of processing natural language.

Finally, using the sentiment scores for sentiments regarding particular product or service with the user’s information, it could successfully profile the products, analyze trends and forecasting. So, as overall, the system is capable of saying that how a set of people of a particular age range, a particular area with a particular profession think about a particular product or service and how it will change it the future which are most useful information when it comes to business world.

ACKNOWLEDGMENT

We would like to thank the staff members at the Faculty of Information Technology for providing valuable insights and comments and helping us in many ways.

REFERENCES

[1] David Osimo and Francesco Mureddu, "Research challenge on Opinion Mining and Sentiment Analysis"

[2] Maura Conway, Lisa McInerney, Neil O’Hare, Alan F. Smeaton, Adam Berminghan, "Combining Social Network Analysis and Sentiment to Explore the Potential for Online Radicalisation," *Centre for Sensor Web Technologies and School of Law and Government*.

[3] Adobe® SocialAnalytics, powered by Omniture®.

[4] Brandwatch. [Online].http://www.brandwatch.com/

[5] Sentiment140.[Online]. http://www.sentiment140.com

[6] Fabrizio S. Andrea E., "Determining the Semantic Orientation of Terms through," October 31– November 5 2005.

[7] H. Chen, "Knowledge Management Systems: A Text Mining Approach", http://ai.bpa.arizona.edu/go/download/chenKMSi.pdf, 2001

[8] I. Nonaka, and H. Takeuchi. The knowledge-creating company, Oxford University Press, Oxford, UK, 1995.

[9] M. Polayni. The tacit dimension. The University of Chicago Press, 1966.

[10] Learning to Classify Test [Online]. http://nltk.googlecode.com/svn/trunk/doc/book/ch06.html#document-classify-all-words