Sentiments Extraction and Label Assignment on Twitter Dataset

Durgesh M. Sharma, Mirza Moiz Baig
J. D College of Engineering and Management
Nagpur, India
durgesh_sharma54@yahoo.com, mirzammb@gmail.com

ABSTRACT: *Sentiment analysis is a process to determine the view on targeted keyword from the internet which is unknown for the users. Users are the customers who share their sentiments on Social Network sites and it make a valuable platform for tracking and analyzing the public sentiments. Tracking and analyzing the sentiments need to extract the sentiments from the internet and there is one site i.e. Twitter. It is a micro blogging site that allows people to share their opinions in 140 characters only. So, mining can easily be done. In this work, we use Twitter for extracting tweets using TF*IDF algorithm and without using any special software we would apply the labeling technique for finding the polarity on tweets viz. positive, negative or neutral.*

Keywords: Extraction, Label assignment, Twitter, Sentiment Analysis, Micro blogging

Received: 25 February 2015, Revised 28 March 2015, Accepted 4 April 2015

1. Introduction

Sentiment analysis is a process to determine the targeted keyword from the internet which is unknown for the users. With the rapid growth of internet, extremely large amount of product reviews are rapidly upload on it [1].Among the myriad types of information available, one useful type is the sentiments or opinions people express towards a subject.

A subject is either a topic of interest or a feature of the topic. For example, knowing the own reputation or their competitors' products or brands are valuable for product development, marketing and consumer relationship management. Traditionally, companies conduct consumer surveys for this purpose.

Well-designed surveys can provide quality estimations, they can be costly especially if a large volume of survey data is gathered [2]. As well as customers can obtain these reviews and can decide to purchase for his/her desire product. So, to give reviews on the products and get feedback on these products, social network sites are the best resource for tracking and analysing public sentiment.

Such tracking and analyzing can provide critical information for decision making in various domains. So, it has attracted attention in both academia and industry [1]. Among various sites there is one site i.e. Twitter, which facilitates to share the views in just 140 characters and makes great market research tool for research.

Many brands are scared to come on social media platforms like Twitter since they feel that they don't have control over what's public comment. But, if brands take each and every user's feedback seriously and work on them, they will know a lot about themselves than any market research company could tell them. For instances, if a lot of people complaint about coffee shop "It's not good", it means time to trained his staff to improve service quality. Negative feedback provides great market research insights, and instigates companies to improve service and efficiency [3].

The public sentiment analysis method has spread and making applications and developments came into existence in this area and now its main target is to make computer able to identify and create emotions like human being [1].For doing this, we are applying the TF*IDF method which extracts10,000latest tweets only which is limited by our source code can be increased as per our need and then pre-process it by translating all slangs words (e.g., lol, omg).

Removing URLs because a lot of users share URLs in their tweets which complicates the sentiment analysis process, removing all stop words(e.g., of, the, because, above) and finally find the polarity from extracted and pre-processed tweets by applying the extensive method.

The extensive method which would apply because we do not need any additional software for putting the polarity but other research scholars need of it and our research work is to there.

2. Related Work

There are various papers published in the field of Sentiment Analysis and we included some of them to give idea in this field. Sentiment analysis on Twitter is the upcoming trend with researchers recognizing the scientific trials and its potential applications. The challenges unique to this problem area are largely attributed to the dominantly informal tone of the micro blogging [4]. Rationale the use of micro blogging and more particularly Twitter as a corpus for sentiment analysis. They cited [5]:

- Micro blogging platforms are used by different people to express their opinion about different topics, thus it is a valuable source of people's opinions.
- Twitter contains an enormous number of text posts and it grows every day. The collected corpus can be arbitrarily large.
- Twitter's audience varies from regular users to celebrities, company representatives, politicians, and even country presidents. Therefore, it is possible to collect text posts of users from different social and interests groups.
- Twitter's audience is represented by users from many countries.

They showed how to automatically collect a corpus for sentiment analysis and opinion mining purposes. They performed linguistic analysis of the collected corpus and explain discovered phenomena. Using the corpus, they built a sentiment classifier that is able to determine positive, negative and neutral sentiments for a document. The polarity of a word can be identified by studying the occurrence frequency of the word in a large annotated corpus of texts. If the word occurs more frequently among positive texts, then its polarity is positive. If it occurs more frequently among negative texts, then its polarity is negative. If it has equal frequencies, then it is a neutral word [6]. The Semantic approach gives sentiment values directly and relies on different principles for computing the similarity between words.

This principle gives similar sentiment values to semantically close words. WordNet for example provides different kinds of semantic relationships between words used to calculate sentiment polarities. WordNet could be used for obtaining a list of sentiment words by iteratively expanding the initial set with synonyms and antonyms and then determining the sentiment polarity for an unknown word by the relative count of positive and negative synonyms of this word [7].

The Semantic Approach is used in many applications to build a lexicon model for the description of verbs, nouns and adjectives to be used in SA(Sentiment Analysis) as the work presented by Maks and Vossen. Their model described the detailed subjectivity relations among the actors in a sentence expressing separate attitudes for each actor. These subjectivity relations are labelled with information concerning both the identity of the attitude holder and the orientation (positive vs. negative) of the attitude. Their model included a categorization into semantic categories relevant to SA. It provided means for the identification of the attitude holder, the polarity of the attitude and also the description of the emotions and sentiments of the different actors involved in the text. They used Dutch WordNet in their work. Their results showed that the speaker's subjectivity and sometimes the actor's subjectivity can be reliably identified [8].

Haddia, Liua and Yong explored the role of text pre-processing in sentiment analysis, and report on experimental results that demonstrate that with appropriate feature selection and representation, sentiment analysis accuracies using support vector machines (SVM) in this area may be significantly improved [9].

3. Sentiments/Opinion Extraction and Label Assignment

The goal of opinion extraction is to detect where in documents opinions are embedded. Opinions are hidden in words, sentences and documents. An opinion sentence is the smallest complete semantic unit from which opinions can be extracted. The sentiment words, the opinion holders, and the contextual information should be considered as clues when extracting opinion sentences and determining their tendencies [10].

For carrying out the Sentiment/Opinion Analysis and finding the variations on sentiments on the context of given targeted keyword, we need to go through following phases as depicting in below Figure 1.

1. Extraction and Pre-processing the Tweets.
2. Sentiment Label Assignment.
3. Sentiment Variation Tracking.

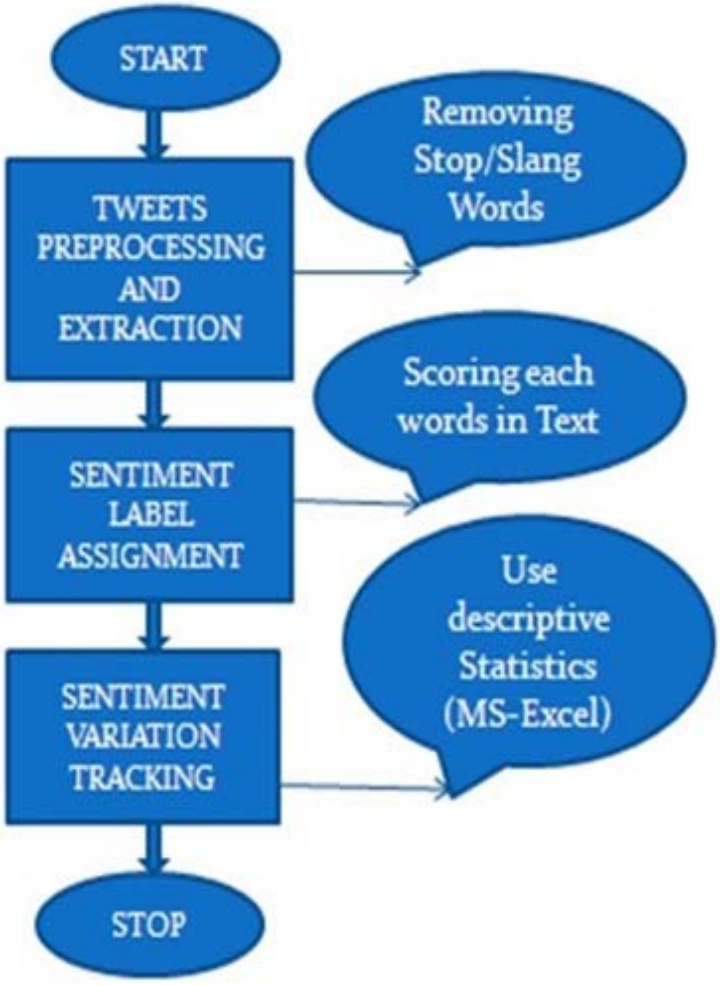

Figure 1. Method for the prediction of Sentiment Analysis and Variation Tracking

3.1 Extraction and Pre-Processing the Tweets

There are the following algorithms for keyword extraction as follows:

1. Word Frequency Analysis also known as TF*IDF (Term Frequency * Inverse Document Frequency).

2. Word Co-Occurrence Relationships

3. Frequency-Based Single Document Keyword Extraction

4. Content-Sensitive Single Document Keyword Extraction

5. Keyword Extraction Using Lexical Chains

6. Key phrase Extraction Using Bayes Classifier

Consider term *t* and document *d* 2 D, where t appears in *n* of *N* documents in *D*. The *TF*IDF* function is of the form:

$$TF - IDF(t, d, n, N) = TF(t, d) \times IDF(n, N) \quad (1)$$

There are many possible *TF* and *IDF* functions. Practically, nearly any function could be used for the *TF* and *IDF*. Regularly-used functions include[11]:

$$TF(t, d) = 1 \text{ if } t \in d$$
$$0 \; else \quad (2)$$

$$TF(t, d) = \sum 1 \; if \; word = t$$
$$0 \; else$$
$$\text{word} \in d \quad (3)$$

Additionally, the term frequency may be normalized to some range. This is then combined with the *IDF* function. Examples of possible IDF functions include:

$$IDF(n, N) = log(N/n) \quad (4)$$

$$IDF(n, N) = log(N - n/n) \quad (5)$$

Thus, a possible resulting TF-IDF function could be:

$$TF * IDF(t, d, n, N) = \sum 1 \; if \; word = t \quad else \times log \; N - n/n \quad word \in d$$

When the *TF * IDF* function is run against all terms in all documents in the document corpus, the words can be ranked by their scores. A higher *TF * IDF* score indicates that a word is both important to the document, as well as relatively uncommon across the document corpus. This is often interpreted to mean that the word is significant to the document, and could be used to accurately summarize the document.*TF * IDF* provides a good heuristic for determining likely candidate keywords, and it (as well as various modifications of it) has been shown to be effective after several decades of research. Several different methods of keyword extraction have been developed since *TF * IDF* was first published in 1972, and many of these newer methods still rely on some of the same theoretic backing as *TF*IDF*. Due to its effectiveness and simplicity, it remains in common use today.

S. Robertson concluded that *TF * IDF* is one of the best-known and most commonly used keyword extraction algorithms currently in use when a document corpus is available[13]. Several newer methods adapt *TF * IDF* for use as part of their process, and many others rely on the same fundamental concept as *TF * IDF*. Nearly all keyword extraction algorithms which make use of a document corpus depend on a weighted function which balances some measure of term or phrase appearance within a document (frequency, location within document, co-occurrence with other words) with some similar measure from the corpus [14].

So, to extract tweets related to the target, Tan et al and we gone through the whole dataset and extract all the tweets which contain the targeted keywords. Compared with regular text documents, tweets are generally less formal and often written in an ad hoc manner. Sentiment analysis tools applied on raw tweets often achieve very poor performance in most cases. Therefore, pre-processing techniques on tweets are necessary for obtaining satisfactory results on sentiment analysis[15]

(1) Slang words translation: Tweets often contain a lot of slang words (e.g., lol, omg). These words are usually important for sentiment analysis, but may not be included in sentiment lexicons. Since the sentiment analysis tool they gone to use based on sentiment lexicon, they converted these slang words into their standard forms using the Internet Slang Word Dictionary and then added them to the tweets[16]

(2) Non-English tweets filtering: Since the sentiment analysis tools to be used only work for English texts, they removed all non-English tweets in advance. A tweet considered as non-English if more than 20 per cent of its words(after slang words translation) do not appear in the GNUA spell English Dictionary[15].

(3) URL removal: A lot of users include URLs in their tweets. These URLs complicate the sentiment analysis process. They decided to remove them from tweets[15].

For analysing the sentiments, we are applying the TF*IDF method which extracts 10,000 latest tweets only which is limited by the source code and can be increased anytime as per our need in the analysis.

So, enter any keyword let say on our Prime Minister Sri Narendra Modi for Chief Minister Election in Delhi Election, entering the keyword “Modi”. After entering the targeted keyword, we get following tweets.

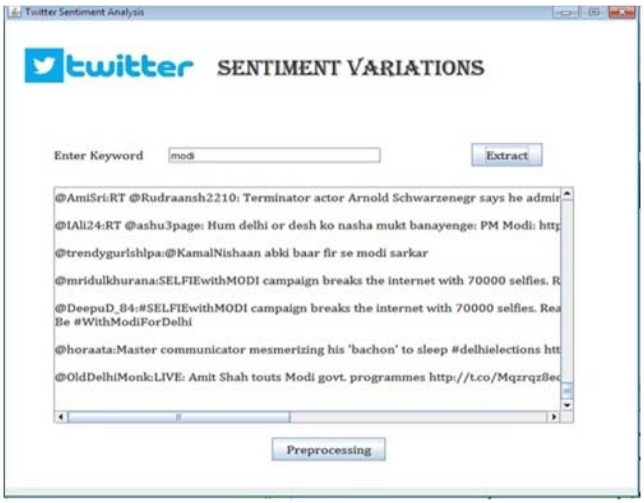

Figure 2. Extracting the tweets from the Twitter

Click on Pre-processing button to remove all stop words, slang words and URLs, for obtaining satisfactory results on the sentiment analysis process and it can concentrate on the targeted keyword only.

Figure 3. Removing all stop, slang and URL words from tweets

Above has done through following lines of code as:

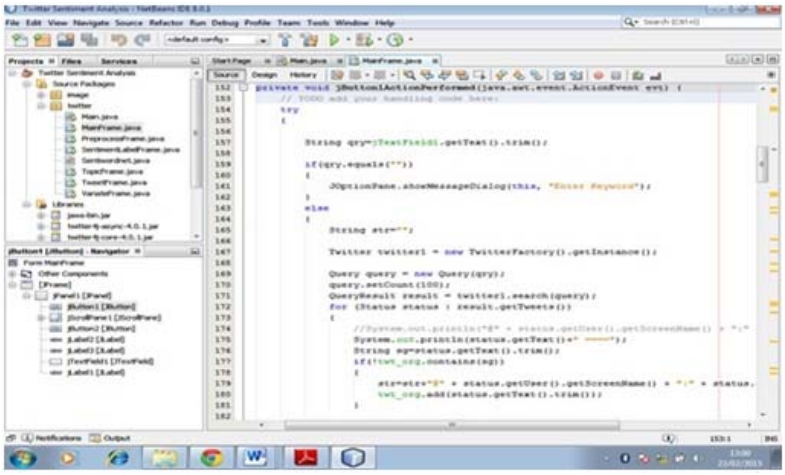

Figure 4. Code for removing the all stop, slang words and URLs from tweets

3.2 Sentiment Label Assignment

Sentiment is a view or opinion of someone on targeted keyword. The area of Sentiment Analysis intends to comprehend these opinions and distribute them into the categories viz. positive, negative, neutral reviews likewise. This task is called Sentiment Label Assignment. Actually, behind the graphical user interface, the below formula calculates the value and the code check the result value in the Positive.txt, Negative.txt along with the SentiWordNet_3.0.0.txt file for assigning sentiments' label. The formula given by [17].

Objective Score = 1 - (Positive Score + Negative Score)

So, after getting Pre-processed tweets on Mr. Modi then click on Sentiment Label Assignment to observe the type of polarity as:

Figure 5. Labelling the Sentiments

Twitter Dataset

Our proposed model applied on a Twitter dataset to label on extracted tweets. The dataset is obtained from the Twitter website online which extracts latest 10,000 tweets and applying the labelling technique on it as:

Objective Score = 1 - (Positive Score + Negative Score)

Above step has done through following lines of code as:

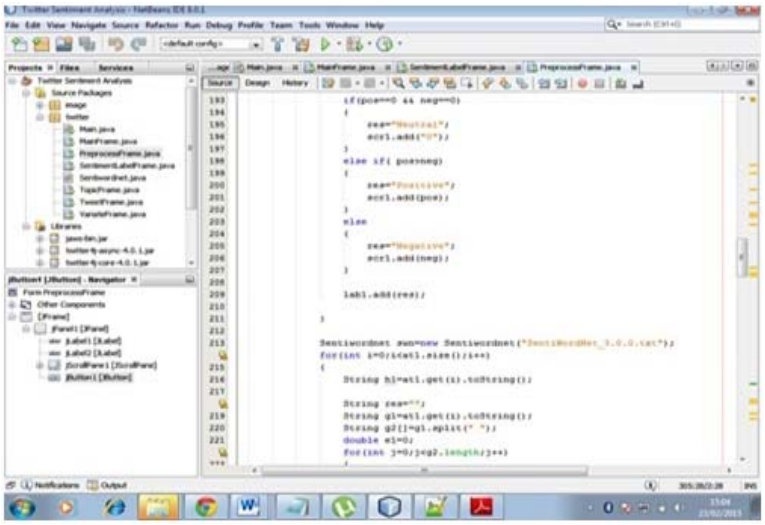

Figure 6. Code for Labelling the Sentiment

Each term in SentiWordNet is associated with numerical scores for positive and negative sentiment information. This database is built upon a subset of words whose priori polarity is known, such as the words “good” and “bad”. It is extended by an iterative process using relationships between words extracted from WordNet database. Each word can be associated with three different scores, positive, negative and objective. SentiWordNet provides positive and negative scores directly. Objective score calculated using above equation[17].

Experiments and Results

We are using the TF*IDF (Term Frequency-Inverse Document Frequency) method because it is the best method for extracting the tweets from the Twitter and we are extracting 10,000 latest tweets only and can be increased the limit from the source code. Then we applied the pre-processing technique where it removes all slang words, stop words, URLs, etc. as already discussed in the above section.

To assign labels on each tweet, many research scholars use additional software but we are labelling autonomously using the code which calls some files viz. Positive.txt, Negative.txt and SentiWordNet_3.0.0.txt along with above discussed formula. These files creates the CSV(Comma Separated Value) files and scores on each words The code finds the words from these files and labels on the tweets as we showed output in Figure 5.

4. Conclusion

It is a process to determine the targeted keyword from the internet which is unknown for the users. Users are the customers who share their sentiments on Social Network sites and it make a valuable platform for tracking and analyzing the public sentiments and that’s why we used Twitter for our research work. In this paper, we showed that for labeling the sentiments on extracted tweets, we have not used any additional software but many research scholars used additional software for it. The above discussed two steps i.e. Tweets extraction and Sentiment label assignment are important for further steps in research work like for reviewing the sentiments on targeted keyword, for finding the reason behind sentiment variations, for improving the next coming product in the market, etc.

References

[1] Sharma, D., Baig, M. (2015). Sentiment Analysis on Social Networking: A Literature Review. *International Journal on Recent and Innovation Trends in Computing and Communication*, 3 (2) 22-27.

[2] Jeonghee, Y., Tetsuya, N., Razvan, B., Wayne, N. (2003). Sentiment analyzer: extracting sentiments about a given topic using

natural language processing techniques. Third IEEE International Conference on *Data Mining* (ICDM), 427 – 434.

[3] http://blog.digitalinsights.in/10-reasons-why-businesses-should-choose-twitter-over facebook/0583787.html (Accessed on 5th Feb 2015).

[4] Kumar, A., Sebastian, T M. (2012). Sentiment analysis on twitter. *IJCSI International Journal of Computer Science Issues*, 9 (3) 372-378.

[5] Pak, A., Paroubek. P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. *In*: Proceedings of the Seventh Conference on *International Language Resources and Evaluation*, 1320–1326.

[6] Read, J., Carroll, J., (2009). Weakly supervised techniques for domain independent sentiment classification. *In*: Proceeding of the 1st International CIKM workshop on Topic-*Sentiment Analysis for Mass Opinion*, 45–52.

[7] Kim, S.M., Hovy, E. (2004). Determining the sentiment of opinions. *In*:Proceedings of COLING-04 20th International Conference on *Computational Linguistics*, 1367–1373.

[8] Maks, I., Vossen, P. (2012). A lexicon model for deep sentiment analysis and opinion mining applications. *Decision Support Systems*, 53 (4) 680-688.

[9] Haddia, E., Liua, X., Yong, S. (2013). The role of text pre-processing in sentiment analysis. *In*: Elsevier Procedia Computer Science, 26 – 32.

[10] Ku, L. W., Liang, Y. T., Chen, H. H. (2006). Opinion extraction, summarization and tracking in news and blog corpora. AAAI *Symposium on Computational Approaches to Analysing Weblogs* (AAAI-CAAW), 100–107.

[11] Salton,G., Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24 (5) 513–523.

[12] Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28 (1) 11–21.

[13] Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for idf. *Journal of Documentation*, 60 (5) 503–520.

[14] Lott, B. (2012). Survey of Keyword Extraction Techniques.

[15] Tan, S., Li, Y., Sun, H., Guan, Z., Bu J., Chen, C., He, X., Yan, X. (2014), Interpreting the public sentiment variations on twitter, *IEEE Transactions On Knowledge And Data Engineering*, 26 (5) 1158-1170.

[16] Thelwall, M., Buckley, K., Paltoglou, G., Cai, D. and Kappas, A. (2010). Sentiment strength detection in short informal text. *J. Amer. Soc. Information Science Technology*, 61 (12) 2544–2558.

[17] Miller, G. A., Beckwith, R., Fellbaum,C., Gross, D., Miller, K. J. (1990). Introduction to WordNet: an on-line lexical database. *International Journal of Lexicography*, 235–244.