IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555
Vol.6, No3, May-June 2016

Lexicon Based Approach to Enhance the Accuracy of Sentiment Analysis on Tweets

G. Vaitheeswaran
Research Scholar,
Department of Computer Science,
St. Joseph's College (Autonomous),
Tiruchirappalli, India.

L. Arockiam
Associate Professor,
Department of Computer Science,
St. Joseph's College (Autonomous),
Tiruchirappalli, India.

***Abstract*—In day-today life, social media websites are serving as a powerful platform for sharing opinions which contains rich source of information on different aspects of zillion users'. Information can be extracted in the form of sentiment polarity from the massive amount of unstructured/structured data by the analytical process, known as Sentiment Analysis. Twitter is a microblogging social media website contains rich source of information to carry-out sentiment analysis. A new approach Senti_Lexi has been proposed to provide better accuracy. The concept of the proposed Senti_Lexi based approach is to evaluate the sentiment knowledge on tweets using lexicon based approach. The emoticons or smileys change the polarity of the sentences. To enhance accuracy the emoticons are used for polarity calculations. The existing sentiment dictionary was built with the unigrams pattern. On the other hand, it is difficult to find polarity of the sentiment word with bigram patterns that has a negation word. To solve these issues, the three new dictionaries are built for the proposed work. They are as follows: Emoticon dictionary, Sentiment dictionary and Negation dictionary. Including the unigrams and bigrams along with emoticons, this research work has gained better accuracy.**

Keywords- sentiment analysis, lexicon based approach, emoticons and negation words.

I. INTRODUCTION

Sentiment analysis (SA) is the field of study that analyses peoples' opinions, sentiments, evaluations, appraisals, attitudes and emotions towards entities such as products, services, organizations, individuals, issues, events, topics and their attributes. It is also called as *opinion mining*, *opinion extraction*, *sentiment mining*, *subjectivity analysis*, *affect analysis*, *emotion analysis*, *review mining*, etc. However, they all are under the umbrella of sentiment analysis or opinion mining.

Bing Liu [1] presented different tasks and works, published in SA and opinion mining. Major tasks listed are subjectivity and sentiment classification, aspect-based SA, sentiment lexicon generation, opinion summarization, analysis of comparative opinions, opinion search and retrieval, opinion spam detection and quality of reviews.

In the sentiment analysis process, lexicon based approach plays a major role to analyze the domain-dependent data sets. To evaluate the accuracy of the proposed approach the real-time tweets are used. A usual tweet holds images, audios, videos, url, word variations, emoticons, hash tags, negation texts, etc., This creates a problem to analyze the polarity of words. Measuring the depth of the sentiment, which mostly rely on the negation word, is one of the major issues and challenges involved in the process of sentiment analysis. The proposed approach mainly focuses on emoticons and negation words. The Senti_Lexi approach is used to classify the tweets as positive, negative or neutral using the emoticon and the negation sentiment orientation of the words.

This article is organized as follows: In section background, a brief introduction about emoticons and negation words are discussed. A brief review towards some related works on sentiment analysis using big data analytics are discussed in the next section. The objective of the proposed approach, required methodology diagram and the proposed Senti_Lexi approach are explained respectively in the successive sections. Finally, the obtained results are discussed and summarized in the last two sections.

II. BACKGROUND STUDY

A. Emoticons

The emoticons are ASCII art, and are called as smileys. The emoticons were originated in 1982, and today the field has reached a major support in various aspects of computer-mediated communication (CMC). In simple term, it adds emotional flavor to the plain text, since it has some sentiment associated with it. The emoticons are seen as social, emotional suppliers to the CMC [2]. The table 1 represents the effects of adding the emoticons in the plain text. As an example [3], imagine the simple phrase *"I'm going to study"*. The addition of an emoticon suffix has greatly changed the context of the message, refer table 1.

TABLE 1 EFFECT OF EMOTICONS TO THE PLAIN TEXT

S. No.	Text	Meaning
1.	I'm going to study :-)	Happy smile
2.	I'm going to study :-(	Sad frown
3.	I'm going to study :-||	Angry
4.	I'm going to study ;-)	Wink

5.	I'm going to study :-D	Deep laugh
6.	I'm going to study :-/	Uneasy
7.	I'm going to study :-P	Tongue out

B. *Negations*

Negation refers to the process of converting the sentiment text from positive to negative or negative to positive by using special words, such as 'no', 'not', 'do not'. The examples of some negation words are presented in the following table 2.

TABLE 2 EXAMPLES OF NEGATION WORDS

Barely	Neither	Nor	Not	Either
Never	Ever	Not any	No	Hardly
Rarely	Scarely	Seldom	Could not	Lack
Dare not	Should not	Does not	Was not	Will not
Have not	Without	Cannot	Do not	Did not

Handling negation words in the sentiment analysis is a important process. The whole sentiment of the text may be changed by the use of negation [4]. The simplest approach to handle negation is to revert the polarity of all words that are found between the negation and the phrase or sentiment words. For instance, in the text 'I do not want to enjoy', the whole phrase 'want to enjoy' will be reverted.

III. RELATED WORKS

Suresh et al. [5] had highlighted the web as an excellent source for assembling consumer opinions such as customer reviews of products, forums, discussion groups, and blogs. This paper mainly focused on online customer reviews of products and made two contributions. First, it proposed a framework for analyzing and comparing consumer opinions of competing products in map and reduced environment for better analysis. Second, a new lexicon-based technique had been proposed to extract neutral reviews and restrict them from being categorized under positive or negative. Experimental results had proven the effectiveness of the proposed technique and its ability to defeat the existing method significantly.

Geeta et al. [6] adjudged that, it is not necessary to use external dictionaries or any other lexicons of polarized words to find the sentiment polarity in tweets. A training data set was automatically generated by referring to the sentiment present in tweets containing emoticons (smileys). It was able to map all common expressions with new words, slangs, and errors.

Ilkyu Ha et al. [7] had proposed other processes of sentiment analysis on hadoop framework to enable parallel process of data. The author used HDFS for storage and MapReduce function for sentiment analysis. It had reduced the time through parallel processing.

Chetan Kaushik et al. [8] had found a technique to efficiently perform sentiment analysis on big data. In this article, sentiment analysis was performed on a large data set of tweets using Hadoop and the performance of the technique was measured in the form of speed. The author had considered the negation and eliminated the emoticons in the preprocessing phase. This technique produced 73.5% accuracy and the experimental result had proved the efficiency of the proposed technique in handling big sentiment data sets.

IV. OBJECTIVE

From the above related works, most of the works had been carried out using the emoticon and negation words separately to find the polarity of tweets. This motivated to build a new model, based on combining emoticon and negation words identification. The objectives of the proposed approach are:

- To label positive, negative and neutral tweets.
- To enhance the precision, recall, f-score and accuracy for positive, negative and neutral tweets.

V. METHODOLOGY DIAGRAM

The overall process of the preprocessing and proposed approach has been divided into three phases. Figure 1 exhibits the methodology diagram of the three phases. They are explained below.

Phase 1: The collected tweets are preprocessed and stored in the desired format (.txt and .csv). The preprocessing step includes stopwords removal, url removal, audio and video removals, POS tagging, stemming, and lemmatization. These processes are briefly explained in the following subsections.

Phase 2: Senti_Lexi approach has been proposed for emoticon handling and negation words; and are briefly explained in the proposed work section. A mathematical equation has been framed for Sentiment Score Computation (SSC) to classify the tweets as positive, negative and neutral.

Phase 3: The positive, negative and neutral tweets are classified and discussed in the results and discussions section of this article. The commonly used polarity measures such as precision, recall, accuracy and F-score are calculated to measure the accuracy of the Senti_Lexi approach.

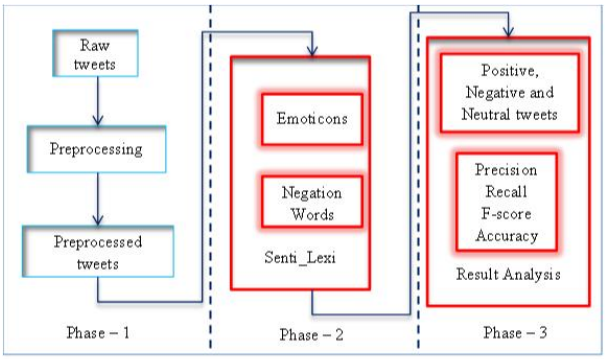

Fig 1 Methodology diagram

A. *Raw tweet*

The raw tweet contains many attributes such as created date & time, user name, text of tweet, location, retweet status, retweet count, etc. For the proposed work, "created_at", "text", "retweet_count" are considered and extracted through the preprocessing step.

B. Preprocessing

The raw data having polarity is highly susceptible due to inconsistency and redundancy. The quality of the data affects the results and therefore in order to improve the quality, the raw data is pre-processed. In this preprocessing step, it removes the repeated words to improve the efficiency of the tweets.

People often use repeating characters while using colloquial language, like "Im in loooove", "We won, yaaayyyyy!" For example,

Raw tweet: "@Iphone is Beauuuutifull"

After preprocessing: "iphone is beautiful".

C. Preprocessed tweets

The preprocessed tweets contains text with emoticons (ASCII character), tweet id, retweet count, date and time.

VI. PROPOSED APPROACH

The concept of the proposed Senti_Lexi approach is to evaluate the sentiment knowledge on big data using lexicon based approach. Tweets displaying negative or positive sentiments are labeled accordingly. If there is no sentiment, then the tweet is marked as neutral.

In the existing work Algo_Sentical approach [8], the emoticons were not considered and eliminated in the preprocessing stage. The emoticons or smileys change the polarity of the sentences.

In this proposed Senti_Lexi approach the emoticons are added to produce better accuracy. The existing sentiment dictionary was built with the unigrams pattern. On the other hand, it is difficult to find the polarity of the sentiment word with bigram patterns that has a negation word. To solve these issues the three new dictionaries are built for the proposed work. They are as follows:

- *Emoticon dictionary*: The positive and negative emoticons are collected manually and assigned to 1 and -1 respectively. The following table 3 represents the example of the positive, negative and neutral emoticons.

TABLE 3 EXAMPLES OF EMOTICONS

Positive Emoticons	*Negative Emoticons*	*Neutral Emoticons*
:)	:(	:/
:-)	:-(	:P
=]	=(	;-)
:D	=/	:0
;)	:@	:|
:0	>:)	:S

- *Sentiment dictionary*: To enrich the sentiment dictionaries the popular existing dictionaries such as Bing Liu, SentiWordNet, and NRC emotions are collected, and combined. Later, the new dictionary is revised to eliminate the repeated words.
- *Negation dictionary*: The negation words are collected manually and assigned with (-1).

Figure 2 shows the pseudo-code for the Senti_Lexi approach and its related procedures and notations. The preprocessed tweets are considered as 'T' and each tweet is considered as 't'. This method is based on the sentence level calculation. Each word in the tweets are considered as 'w'.

Initially the polarity of negation is calculated using the proposed formula, and is discussed in the handling section. The proposed method calculates the polarity of the sentiment based on the following features:

- *Emotions only* : Find the polarity for the sentiment words (unigram model).
- *Emotions with Emoticons and Negation words* : Find the polarity for the sentiment words along with the emoticons and negated words.

Approach: Senti_Lexi.

Find the positive, negative and neutral tweets based on lexicon classification.

Input: T, preprocessed tweets

Output:

- Labeled positive, negative and neutral tweets.
- Accuracy of overall sentiment classification.

Method:

```
nw = ["not", "didn't", …]
happy = [":)", ":-)", ":-D", … ]
sad = [":(", ":-(", … ]
pos = positive.txt
neg = neg.txt
for t in T
for w in t
    if(prev_word in nw && w in neg)
    (-1) * (w)      /* polarity of negative word is -1 */
    return positive
    elseif(prev_word in x && w in pos)
    (-1) * (w)      /* polarity of positive word is 1 */
    return negative
    else
    return neutral
                if(w in happy)
                        return positive
                elseif(w in sad)
                        return negative
                else
                        return neutral
                if(w in pos)
                        return positive
                elseif(w in neg)
                        return negative
                else
                        return neutral
        calculate sentence level polarity
        end
print positive, negative and neutral tweets
end
```

Notations:

T as collected tweets
t as each tweet
w as each word in t

nw = list of negation words
happy = list of positive emoticons
sad = list of negative emoticons
pos = list of positive words text file
neg = list of negative words text file

Fig 2 Senti_Lexi Approach

A. *Emoticons*

If a tweet contains only positive emoticons and no negative emoticons, it is classified as positive. If a tweet contains only negative emoticons and no positive emoticons, it is classified as negative.

B. *Negation Handling*

Typically, tweets are very short text messages. The presence of single negation word changes polarity of the sentences. The list of negation words is obtained from the project available on the Github [9].

For eg. "*Raja is not a good boy. He is not bad than Ramu*". Assigning the usual unigram model for the above example, resulted in negative statement. Since the polarity value for not is (-1), good is (1) and bad is (-1). Calculating the sentence polarity,

Not (-1) + good (1) + not (-1) + bad (-1) = -2.

Thus, applying the unigram model for negation handling resulted in negative statement. To overcome the above described problem a new calculation method is proposed for handling the negation words. For this negation handling problem the default value of (-1) is assigned to the negation word list.

if(prev_word in negation_dict && current_word in senti_dict) then multiply the value of the current word with -1. Where negation_dict contains the list of negation words and senti_dict contains the positive and negative words.

The sentiment polarity for the negation word is calculated using the following equation,

$$(-1) * (w_{xy}) \ldots \text{Eq. (1)}$$

Where $w_x = 1$ if it is positive word and,

$w_y = -1$ if it is negative word.

For the better understanding of negation handling words, table 4 helps to know its rules.

TABLE 4 RULES OF NEGATION

Presence of Previous word in negation list	Current word	Negation	Polarity
No	Positive	True	Positive
No	Negative	False	Negative
Yes	Positive	False	Negative
Yes	Negative	True	Positive

C. *Positive, negative and neutral tweets*

Equation 2 represents the calculation of polarity of the unigram words and emoticons.

$$\text{Score (P)} = \begin{cases} 1, & w_x > 0 \\ -1, & w_y < 0 \\ 0, & \text{otherwise.} \end{cases} \quad \ldots\ldots\ldots \text{Eq. (2)}$$

- Where P is the polarity, and w_x, and w_y are the positive, and negative words respectively.

D. *Sentiment Score Computation (SSC)*

Based on the sentence-level calculation, the positive, negative, and neutral tweets are calculated for the collected tweets and labeled. The following equation (3) is used to find the positive, negative, and neutral tweets for each sentence.

$$\text{Score (S)} = \sum_{j=0}^{k} w_j \ldots\ldots\ldots \text{Eq. (3)}$$

- where S is the sentence level score, and w_j is the word of each sentence.

The following equation (4) is used to count the overall polarity score for the collected tweets.

$$\text{Score (TS)} = \sum_{i=0}^{n} S_i \ldots\ldots\ldots\ldots\ldots \text{Eq. (4)}$$

- where TS is the total score of the obtained tweets.

If the TS > 0, the chosen topic is positive, if TS < 0 the chosen topic is negative and otherwise the topic is neutral.

The positive, negative and neutral tweets are identified based on SSC and stored separately in the Comma Separated Value (CSV) file. The CSV files are labeled as positive, negative and neutral.

VII. RESULTS AND DISCUSSIONS

From the above proposed Senti_Lexi approach, the obtained results are discussed in this section.

The tweets are obtained using the Python based "tweepy" API. NLTK tool is used for preprocessing the raw tweets and for polarity calculations.

Figure 3 represents the correctly classified tweets from the raw tweets. The X-axis represents the classified tweets and Y-axis represents the count of classified tweets. The correctly classified tweets are analyzed with the proposed Senti_Lexi approach and labeled as positive, negative and neutral tweets based on the positive and negative emoticons and sentiment words of each tweets. Such sentiment words are analyzed using the proposed negation handling concept, discussed in the previous section.

Figure 4 represents the percentage of classified positive, negative and neutral tweets. The positive tweets with percentage of 58, negative tweets with percentage of 19 and neutral tweets with percentage of 23 are obtained.

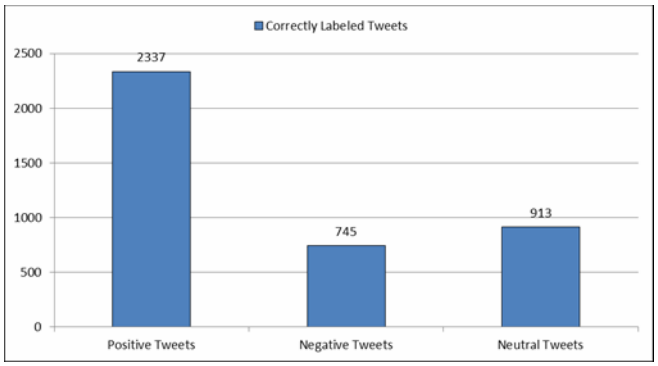

Fig 3 Correctly Labeled Tweets

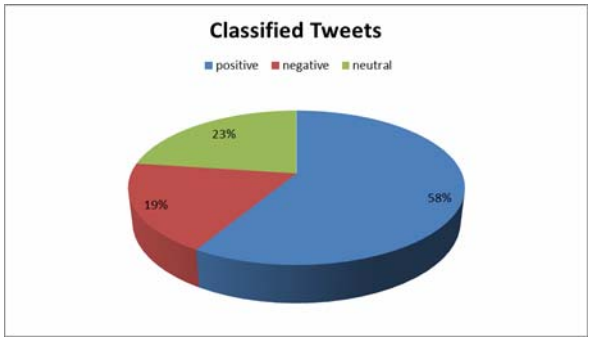

Fig 4 Percentage of Classified Tweets

Figure 5 represents the percentage of classified positive, negative and neutral tweets based on emoticons only. The obtained results are the positive tweets with percentage of 69, negative tweets with percentage of 15 and neutral tweets with percentage of 16.

A comparison made to the above figure 4 along with the figure 5, indicates that there is a contradiction between the analysis of emoticons only and emotions with emoticons and negation words. For better results, it is advisable to consider emoticons along with the words.

Some of the positive emoticons are expressed with the negative statement resulted the bad polarity of the sentences. Hence, creates the context dependent problem between words and emoticons.

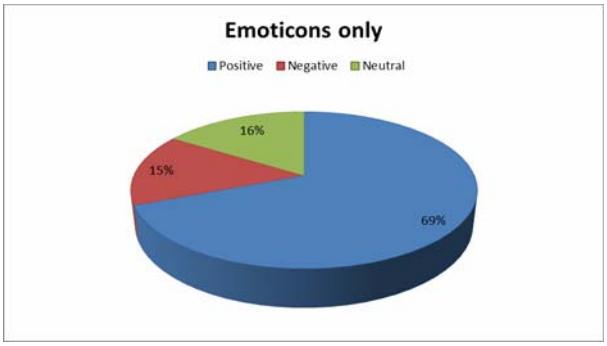

Fig 5 Percentage of Classified Tweets Based on Emoticons only

Figure 6 represents the correctly classified tweets from the raw tweets based on the emoticons only and emotions along with emoticons and negation words. The X-axis represents the features and Y-axis represents the count of tweets. The correctly classified tweets are labeled as positive, negative and neutral tweets with counts are depicted in the same figure.

Figure 7 represents the precision, recall, F-score and accuracy of the collected tweets. The X-axis represents the common measures and Y-axis represents the percentage of common measures. Precision and recall produces an equal result. This shows the efficiency of the proposed approach.

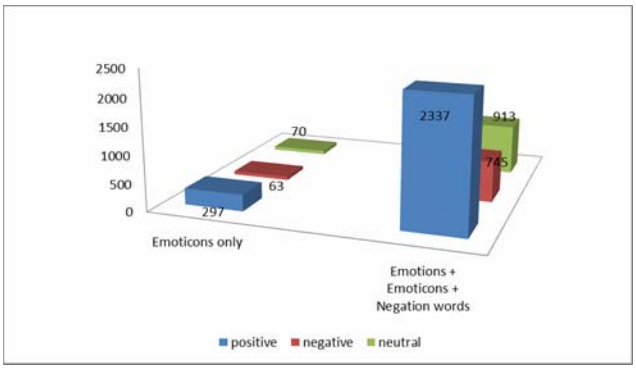

Fig 6 Comparative Results of Classified Tweets based on Features

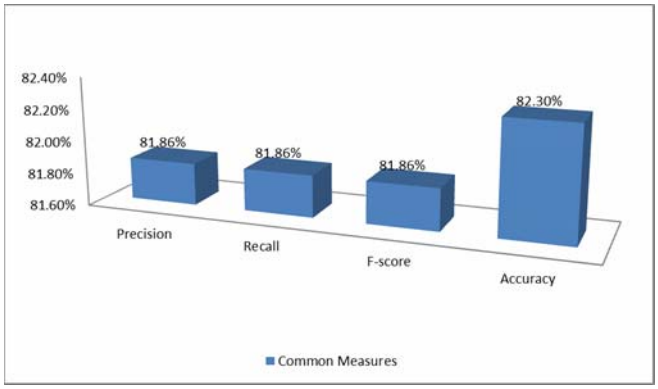

Fig 7 Measurements of Precision, Recall, F-score and Accuracy

Figure 8 represents the comparison of two approaches, namely Algo_Sentical approach and Senti_Lexi approach. The X-axis represents the names of the approaches and the Y-axis represents the percentage of accuracy. The proposed Senti_Lexi approach with a percentage of 82.30 is having higher accuracy than the existing Algo_Sentical approach with a percentage of 73.50. The proposed Senti_Lexi approach is having 8% more accuracy than the existing approach.

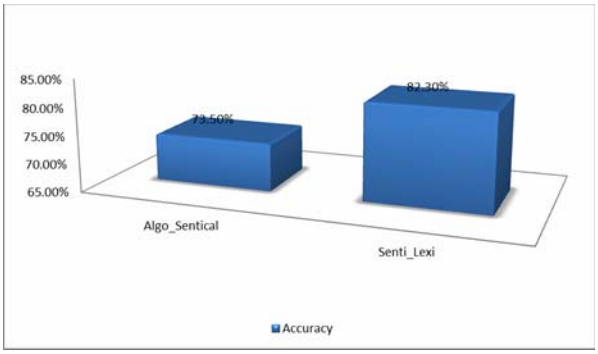

Fig 8 Comparative Results of the Proposed Approach with the Existing Approach

VIII. CONCLUSION

The least amount of research had been carried out in emoticon and negation based analysis, which brings more issues and challenges to the industrialists and academicians. This article has presented a novel lexicon based model for analyzing sentiment analysis on tweets. The methodology diagram discussed in this article delivers a big picture for processing sentiment analysis of social media data. A new approach Senti_Lexi along with emoticon and negation text has been proposed to provide better accuracy than the existing work. In the proposed approach, a 8% increase in accuracy is resulted by adding emoticons. Some of the positive emoticons are expressed along with the negative statement resulted the bad polarity of the sentences. Hence, creates the context dependent problem between words and emoticons. In future, by considering and solving the context dependent problem in sentiment analysis results in better accuracy.

REFERENCES

[1] Bing Liu. *"Sentiment Analysis And Opinion Mining"*, Morgan and Claypool publishers, 2012.

[2] Tanimu Ahmed Jibril and Mardziah Hayati Abdullah. *"Relevance of Emoticons in Computer-Mediated Communication Contexts: An Overview"*, Asian Social Science, Vol. 9, Issue 4, 2013, ISSN: 1911-2017, E-ISSN: 1911-2025.

[3] Internet source as on 18th May 2016, http://www.datagenetics.com/blog/october52012/index.html

[4] Councill I. G., McDonald R., and Velikovich L. *"What's Great and What's Not: Learning to Classify the Scope of Negation for Improved Sentiment Analysis"*, In Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, NeSp-NLP '10, Association for Computational Linguistics, 2010, pp. 51-59.

[5] R. Suresh Ramanujam, J. Nivedha, R. Nancyamala and J. Kokila. *"Sentiment Analysis Using Big Data"*, IEEE 2015, International Conference on Computation of Power, Energy, Information and Communication, Vol 6, 2015.

[6] Geeta.G.Dayalani, Dr. Seema, Prof. B. K. Patil. "*Emoticon-Based Unsupervised Sentiment Classifier for Polarity Analysis in Tweets"*, International Journal Of Engineering Research and General Science, Vol 2, Issue 6, 2014.

[7] Ilkyu Ha, Bonghyun Back and Byoungchul Ahn. *"MapReduce Functions to Analyze Sentiment Information from Social Big Data"*, International Journal of Distributed Sensor Networks, 2015.

[8] Chetan Kaushik and Atul Mishra. *"A Scalable, Lexicon Based Technique For Sentiment Analysis"*, International Journal in Foundations of Computer Science & Technology (IJFCST), Vol 4, Issue 5, 2014.

[9] Internet source as on 18th May 2016, https://github.com/jeffreybreen/twitter-sentiment-analysis-tutorial-201107/blob/master/data/opinion-lexicon-English/negative-words.txt