Article can be accessed online at http://www.publishingindia.com

A Lexicon Based Method for Opinion Mining

Adavi Lakshmi Bhargav*, B Prajna**

Abstract

Nowadays every user wants to know about particular product before buying, movie reviews before watching to confirm whether it is good or bad. The developers are also interested to know about their products or movies based on the user reviews. For that purpose the sentiment analysis is very useful.

The sentiment analysis is very important to know about the product reviews, movie reviews, tweets and so on. Based on that reviews, the user can classify whether it is good or bad (positive opinion or negative opinion). So the sentiment analysis plays an important role in human life.

This paper describes the simple lexicon based approach for classifying the sentence to positive or negative or neutral. Our aim in sentiment analysis is to produce summary of opinion based on product features and reviews. For this process we create a lexicon. The lexicon contains positive, negative, negation, blind negation words and emoticon list. The product reviews, movie reviews, tweets contain word variations, emoticons, hash tags etc. We perform some steps to process the sentence that has hash tags and exaggerated word shortening. After the pre-processing our lexicon based approach tells that the sentence is positive or negative or neutral based on the adjectives present in the sentence.

Keywords: Opinion Mining, Lexicon, Polarity

1. Introduction

Nowadays social networking sites (e.g., facebook, twitter) and online purchasing sites (e.g., e-bay, flipcart) are very popular. Through the social networking sites, people can share the feelings (Kaufmann & Kalita, 2010; Go *et al*., 2009). People mostly depend on the online purchasing sites to purchase any product. In the olden days people take decisions through the friends and relatives to buy a product. But nowadays there is no need to asking others because many tools and sites are there to serve the people. In those sites the users post their opinions based on which the new users can decide whether it is good or bad. The developers are also interested to know about their product. For that purpose the sentiment analysis is taken in consideration. The sentiment analysis is also known opinion mining. The main aim of the sentiment analysis is to find whether a given sentence is positive or negative or neutral (Liu, 2008).

The rest of the paper is organised as follows. In the second section, we can know about related work and different methods about this process. In the third section, we describe processing steps in our implementation method. In the fourth section, we test the given data. In the fifth section, we evaluate final result. The sixth and the last section describes the conclusion and future work.

2. Related Work and Different Methods

The sentiment analysis is mainly used to know whether the sentence or document is positive, negative or neutral. The users post their opinions regarding a particular product in

* Computer Science and Systems Engineering, Andhra University College of Engineering (A), Visakhapatnam, Andhra Pradesh, India. Email: bhargavdv454@gmail.com

** Associate Professor (CS&SE), Andhra University College of Engineering (A), Visakhapatnam, Andhra Pradesh, India. Email: prajna.mail@gmail.com

a sentence or a document manner (Liu, 2008; Pang & Lee, 2008). If we want to know about that product, we must read the sentence or document. The sentence or document has word variations, emoticons, # tag statements, exaggerated words etc. To understand such sentences or documents, user feels some difficulty, as mainly in product reviews and movie reviews, users post some parts to be good and some to be bad. In such situation the buyers are confused about the product. For example, the Samsung phone battery life is good, applications is awesome but price is high, screen resolution is bad. This post may confuse some users. They may think whether the product is good or bad. To overcome such situation, the sentiment analysis (or) opinion mining is very useful to users. Thus sentiment analysis has become a good research area since 2002.

There are mainly 8 sentiment analysis methods (Conclaves & Araújo, 2010). They are

2.1. Emoticon Based Sentiment Analysis

It is a simple method to describe whether a sentence is positive or negative. It is based only on the emoticons present in the sentence. If the emoticon is in positive list, it returns positive polarity, otherwise it returns negative polarity.

2.2. Linguistic Inquiry and Word Count (LIWC)

This method calculates the sentiment based on the dictionary words. The main advantage of LIWC method is that it provides optimised options. That means we can include our own dictionary instead of previous one.

2.3. Sentiwordnet

The sentiwordnet is more popular sentiment analysis method. In this method a lexical dictionary is used (Esuli & Sebastiani, 2006). The dictionary collects adjectives, nouns, etc. based on which it analyses the sentiment. In this method, the rating falls between 0 and 1.

2.4. Senticnet

The senticnet is another method in sentiment analysis. In this method the main aim is to find the polarity of a given sentence in semantic manner. For that purpose this method uses the natural language processing technique.

2.5. SentiStrength

The sentistrength method is mainly depending on the LIWC method. In the LIWC method we have positive and negative word lists, but this method has booster words and weak words along with positive and negative words.

2.6. Sail Ail Sentiment Analyzer

SASA (Sail Ail Sentiment Analyzer) method is based on the machine learning techniques. SASA is an open source tool and it calculates the sentiment by Amazon Mechanical Turk (AMT). There is no comparison of this tool against other methods.

2.7. Happiness Index

Happiness Index is another method in sentiment analysis. It works mainly based on the ANEW (Affective Norms for English Words). The words have scores ranging between 1 and 9. If the rating of words is between 1 to 4, it is considered to be negative sentence. The rating from 5 to 9 is considered to be positive.

2.8. PANAS-t

The PANAS-t method is mainly based on the psychology. It consists of a large set of words with some moods. The range is between -1.00 to +1.00. Based on those moods it calculates the score for the sentence and returns the polarity.

3. Lexicon Based Method

The lexicon based method (Ding *et al*., 2008; Esuli., & Sebastiani, 2006; Taboada *et al*., 2011) is similar to LIWC method. In the lexicon method we initially create a lexicon. The lexicon may be user defined lexicon; that means it is a manually created lexicon. The lexicon contains the several lists which are sentilist, sentinegation list, blind negation list, and emoticon list.

3.1. Creation of Lexicon

3.1.1. Sentilist

The sentilist consists of common adjectives, that means both positive and negative sentiment words. For example "awesome", "smart" always show positive opinion and "angry", "ugly" always show negative opinion. The range of each word must be between -1 to 1 to indicate the polarity.

3.1.2. Sentinegation List

The sentinegation list consists of negation words (Wiegand *et al*., 2010). The negation words reverse the polarity of a sentence.

E.g. "This is not good." Here, adjective "good" is present, so the sentence may be positive, but "not" is present before "good", so the sentence is negative. For this purpose the negation list will be useful.

3.1.3. Blind Negation List

The blind negation list consists of some words that blindly say that the sentence is of negative polarity. E.g. "The computer needs good CPU". Here "needs" is represented as blind negation word. "Good" actually represents positive polarity, but here the sentence is of negative polarity.

3.1.4. Emoticon List

The emoticon list consists of positive emoticons and negative emoticons (Wikipedia).

Table 1: Dictionary Details

List name	*No. of words present*
• Positive	119
• Negative	103
• Blind negation	4
• Negation	16
• Positive emoticon	12
• Negative emoticon	17

3.2. Processing Steps

A difficult sentence contains large number of variations, hash tags, emoticons, exaggerated words etc. The main aim of this paper is to calculate the sentiment for such sentences. For this process we firstly create a lexicon consisting of sentilist, sentinegation list, blind negation list, and emoticon list.

3.2.1. Emoticon Detection

Firstly, find out whether any emoticon is there in the sentence. If the emoticon is present, then go to the emoticon list (Davidov *et al*., 2010) and search in which list (i.e., positive emoticon list or negative emoticon list) it is present. Then return the polarity accordingly.

3.2.2. Exaggerated Word Shortening

This means some users post their opinion like "goodddd" to express their feeling. It is not a spelling mistake; it is rather the user's excitement. In such situation too, we can give the correct result. For that purpose we use the regular expression to shorten the word like "good".

3.2.3. Hash Tag Detection

Consider the statement like "#iamgoodboy". Here there is no space between the words in # tag (Davidov *et al*., 2010). It is difficult to calculate sentiment analysis. To overcome this problem, we compare the sentence with sentilist and get the words with index and sort words to find the polarity of sentence.

3.3. Sentiment Calculation

If the sentence has emoticon, it first checks the emoticon list and returns the polarity. Otherwise the adjectives are extracted from the sentence along with their ratings (Benamara *et al*., 2007). Then add the adjectives' ratings and find the average. If the average is greater than 0 then it returns a positive polarity for the sentence. If the average is less than 0, it returns the sentence has a negative polarity. If the average is equal to 0 then it returns sentence is neutral.

4. Testing the Given Data

The testing data consists of real time reviews given by the users. Above 50 expressions are presented in the experimental data. Those expressions are divided into positive, negative or blind negation etc. The dictionary is manually created as shown in Table. Our program runs on the testing dataset. We calculate the correct sentiment of the data compared to existing one. We can see the testing data in Table 2.

Table 2: Testing Data

Opinion type	*Expression count*
• Positive	33
• Negative	10
• Neutral	7

4.1.1. Algorithm

For sentiment calculation

Data: movie reviews or product reviews or twitter data

Result: Output: Positive. Negative, Neutral.

Collect the list of positive and negative words with ratings (sentiment list).

Collect the list of negation words (sentiment negation).

Collect the list of blind negation words. Collect the list of positive emoticon list.

Collect the list of negative emoticon list.

If Hash tag is present then

Compare the sentence with positive and negative list.

If word is found then store with index. Based on that index, sort the words and add them to sentiment list.

End

If Blind negation then

Return negativity;

Else

If Sentiment list and sentiment negation list then

For each word in the sentiment list do

If word is at most the distance of 2 from sentiment negation list then Revert the polarity of the word.

End

End

Else

If sentiment negation then

Add the sentiment negation list to the negative sentiment list;

End

End

End

Sentiment sum=0;

For each word in the sentiment list do

Sentiment sum= sentiment sum+ sentiment of word;

End

If ! is present then

Sentiment sum= sentiment sum*2;

End

If emoticon is present then

Find the emoticon and search in emoticon list then return the sentiment

End

If sentiment sum>0 then

Sentiment type=”positive”;

End

If sentiment sum<0 then

Sentiment type=”negative”;

Else

Sentiment type=”neutral”;

End

4.1.2. Result

Our sentiment analysis tool worked very well. Precision and recall measurements are calculated and they are shown in Table 3. Here the recall rates are lower than precision because some bad words like s**t, s**k are generally taken as negative sentiment words. But the sentences like "this s**t is good" gives positive polarity though it actually is a bad sentence. These types of sentences are treated as negative sentences. We can see the result in Table 3.

Table 3: Result

	POSITIVE	*NEGATIVE*
PRECISION	0.9216	0.8756
RECALL	0.7315	0.8012

5. Conclusion

In this paper we explained our system that is used for calculating the sentiment of a given sentence. We explain a lexicon based approach for sentiment analysis with movie reviews and product reviews. We provide practical methods for calculating the sentiment if the sentence has emoticon or hash tags or exaggerated words. The lexicon based method is a simple, reliable and practical approach to sentiment analysis with movie reviews and product reviews. The lexicon method is good based on our lexicon creation.

6. Future Work

We are further planning to implement our algorithm for calculating sentiment of the given sentence when spelling mistakes are present in the sentence.

We are also planning to implement our algorithm for calculating the sentiment based on the semantics also.

References

Benamara, F., Cesarano, C., Picariello, A., Reforgiato, D., & Subrahmanian, V. S. (2007). *Sentiment Analysis: Adjectives and Adverbs are Better than Adjectives Alone*. In Proceedings of International Conference on Weblogs and Social Media.

Conclaves, P., & Araújo, M. (2010). UFMG: Comparing and combining sentiment analysis methods. Journal of Language and Social Psychology, 29(1), 24-54.

Davidov, D., Tsur, O., & Rappoport, A. (2010). *Enhanced Sentiment Learning using Twitter Hash Tags and Smileys*. Proceedings of the 23rd International Conference on Computational Linguistics, (pp. 241-249).

Ding, X., Liu, B., & Yu, P. S. (2008). *A Holistic Lexicon-based Approach to Opinion Mining*. Proceedings of the International Conference on Web Search and Web Data Mining, (pp. 231-240).

Esuli, A., & Sebastiani, F. (2006). Sentiwordnet: *A Publicly Available Lexical Resource for Opinion Mining*. Proceedings of LREC, 6, 417-422.

Go, A., Huang, L., & Bhayani, R. (2009). *Twitter sentiment analysis*. Final Projects from CS224N for Spring 2008/2009 at The Stanford Natural Language Processing Group.

Kaufmann, M., & Kalita, J. (2010). *Syntactic Normalization of Twitter Messages*. International Conference on Natural Language Processing Kharagpur.

Liu, B. (2008). *Sentiment analysis and opinion mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Pang, B. & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1), 1-135.

Taboada, M., Brooke, J., Toloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2), 267-307.

Wiegand, M., Balahur, A., Roth, B., Klakow, D., & Montoyo, A. (2010). *A Survey on the Role of Negation in Sentiment Analysis*. Proceedings of the Workshop on Negation and Speculation in Natural Language Processing.

Wikipedia. List of Emoticons. (2011). Retrieved from http://en.wikipedia.org/wiki/List_of_emoticons.