ISSN:0975-9646

Ajay Siva Santosh Reddy/ (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (5) , 2012,5176 - 5178

Polarity Analysis through Neutralization of Non-Polar Words and Segregation of Polar Words Using Training Data

Ajay Siva Santosh Reddy
St. Francis Institute of Technology
Mumbai, India

ABSTRACT-With the advent of the World Wide Web and its ease of access to the general public, people who used to depend on word of mouth to get an opinion or review of products are now turning to the web for the same. This paper proposes a method to mine the numerous reviews available online and help the consumer derive a non-ambiguous idea about the product. Bayes' algorithm in general and PLSA in particular are used to give a polarity to the review of the product. The major concept being the neutralization of non-polar words along with identifying and segregating the polar words for further comparison with text with the help of training data using latent semantic analysis.

Key words- Opinion, training, polarity, Semantic, probability.

1. INTRODUCTION

Sentiment Analysis is a field of data mining in general and subjective analysis in particular. It is a method defined to identify the opinion expressed by a person in either speech or text. As the name says this method is useful to determine the sentiment expressed by a piece of text, the sentiment in this case means the notion expressed by a person on a product of the performer of the sentiment analysis is trying to gain an opinion about. The overall accepted perception of a task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level — whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral. Advanced, "beyond polarity" sentiment classification looks, for instance, at emotional states such as "angry," "sad," and "happy." Turney and Pang are the individuals credited with the earliest work in this aspect; they used this method on the reviews movies and restaurants. The core aspect of this method is natural language processing which uses various techniques on the day to day language used in writing or speech to generate a polarity index which speaks about the sentiment and opinion expressed by the owner of the text.

Now with the ever increasing popularity of various kinds of social on the internet leading to the users spending a lot of time online also performing activities such as expressing their ideas and opinions in various formats such as blogs, comments, reviews on the sites featuring the product and various other kinds of feedbacks available online. These sentiments can be categorized either into two categories: positive and negative; or into an n-point scale, e.g., very good, good, satisfactory, bad, very bad.

It has been noted that all the existing research on textual processing has been focused on mining and factual information and only a little work has been done on the processing of opinions till now. One main reason for this lack of study on opinions is the fact that there was little opinionated text available before the World Wide Web. Now with the explosive growth of the user generated content on the web in the past few years, individuals who used to consult friends or families and organizations that depended on the opinion polls or surveys to make decisions are now turning to the web. So we can safely say that the web has emotions.

II. SOURCES OF DATA

This section highlights the aspects which answer the query as to from where is the data extracted in order to be subjected to mining and in our case sentiment analysis. It has been observed that most of the opinions on the web are expressed in these four forums.

1) Blogs
2) Review Websites
3) User comments on shopping sites
4) Micro-blogging sites

Blogs:

These are places where a user expresses his opinions without any inhibitions and it has been observed that an increasing number of reviews are appearing on the blogs on a daily basis. Hence blogs form an important source of data and opinions.

Review Websites:

There are a lot of review websites whose primary purpose is to review all the latest and existing products related to the field the website caters to. The products are reviewed in depth along different dimensions and each dimension is separately rated. This helps the user gain a perspective but due to the sheer number of such websites available and the different opinions expressed on them these too form an important source for mining and also a need to mine them arises.

User Comments:

These are the comments which usually follow the catalogue listing on a shopping website or the review of the product on the review website. These usually contain the first hand user experiences of the product and also give an idea about the level of satisfaction an user is experiencing. Since it's the reviews straight from the actual customer who would state his opinions without bias these too form an important source for mining and opinion extraction.

Micro-blogging Sites:

On these social networking site people communicate about every potential area of interest, including, Movies, food, sports, music, etc. It is estimated that there are over 900

social media sites on the internet. There are many popular platforms such Facebook, Twitter, LinkedIn, Google Plus, and YouTube with people spending over 500+ billion minutes per month on these sites. There are nearly 500 million registered Twitter accounts and seventy-seven per cent of internet users read blogs. The majority of the population including children and adults is using social media in some form or another. As of July 2012, the average number of tweets sent per day was 140 million. They are not just viewing the content and they are also socializing and tweeting about like they like or not. Extracting data and applying the mining techniques could result in the generation of the overall idea about the opinionated content available on all these networking sites.

III. APPROACH

Since there is a large amount of data obtained from various with different opinions expressed in each of them there is a need to devise a method to mine them efficiently so that a summary of the overall opinion is obtained. For this purpose we follow the sentiment analysis model (fig.1) which gives us a step by step idea of the process.

Fig.1 Sentiment Analysis Model

The first block in the model is the source of the text we plan to implement the analysis upon, this is followed by data preparation which is the process of extracting and storing data for the process of analysis. Next is the actual process of sentiment analysis which uses training data to classify words as neutral or polar and uses either of these sets words to estimate the sentiment and give the overall polarity of the document or a review submitted to the system. This uses the Bayes' probability function do determine the polarity of words and check the occurrence and frequency of the particular word in the review text which has to be analyzed upon to give the polarity of the document due to the occurrence of that particular word through the process of latent semantic analysis.

The training data contains previously segregated positive and negative reviews and they are fed in to the system. All the words in the positive review are considered positive and all the words in the negative review are considered negative. Equal number of positive and negative reviews with approximately same amount of content is fed into the system. The neutrality of the common words in maintained when it appears in both the reviews. The extent polarity of a word is decided with the frequency of occurrence in the positive or negative review.

When a polar word appears in the test review its frequency is multiplied with the previously obtained polarity from the training data and the summation of the polarities is taken to give a figure which represents the overall polarity which can be positive or negative. The equations below give the polarity calculation of a word and the overall polarity calculation.

$$P^{+}(W) = \frac{O(W) \text{ in } P.R}{(O(W) \text{ in } P.R + O(W) \text{ in } N.R)} \quad \ldots. (1)$$

$$P^{-}(W) = -\frac{O(W) \text{ in } N.R}{(O(W) \text{ in } P.R + O(W) \text{ in } N.R)} \quad \ldots. (2)$$

$$P(W) = P+(W) + P-(W) \quad (3)$$

$$P(\text{Document}) = \sum O(Wi) * P(Wi) \quad . \ (4)$$

Where:-
O(W) is the frequency of occurrence of the word
P+(W) is Positive polarity of a word
P-(W) is Negative polarity of a word
P(W) is Overall polarity of the word
P(D) is Overall polarity of the document
i varies from 0 to the total no. of occurrences of the word

IV. WORKING

POSITIVE TRAINING REVIEW	NEGATIVE TRAINING REVIEW
The movie is great.	The movie is bad.

Polarity of Words

$P+(The)= \frac{1}{2}$	$P-(The)= -\frac{1}{2}$	P(The)=0
$P+(Movie)= \frac{1}{2}$	$P-(Movie)= -\frac{1}{2}$	P(Movie)=0
$P+(is)= \frac{1}{2}$	$P-(is)= -\frac{1}{2}$	P(is)=0
$P+(great)= \frac{1}{1}$	$P-(bad)= -\frac{1}{1}$	P(great)=1 , P(bad)=-1

Sample Review:- Bad Movie

P(Document)= (1*P(Bad))+(1*P(Movie))
= -1
Overall Polarity of the document is Negative.

V. CONCLUSION

This method gives us an effective way of classifying a document as a positive and negative with also giving the extent to which it is positive or negative. Millions of reviews which appear online in various forms, expressing a multitude of opinions can be easily sorted and analyzed as it is humanely impossible to go through all the content available online with respect to a product to make an informed decision owing to the highly contrasting nature of the content.

REFERENCES

[1] Hiroshi, K., Tetsuya, N., & Hideo, W. (2004), Deeper sentiment analysis using machine translation technology, In Proceedings of the 20th international conference on computational linguistics (COLING 2004), August 23 – 27, 2004 (pp. 494–500). Geneva, Switzerland.

[2] Hatzivassiloglou, V. & McKeown, K. R. (1997), predicting the semantic orientation of adjectives, In Proceedings of the 8th conference on european chapter of the association for computational linguistics (pp. 174–181). Madrid, Spain.

3]Kim, S.-M. & Hovy, E. (2004), determining the sentiment of opinions, In Proceedings of the 20th international conference on computational linguistics (COLING 2004), August 23 – 27, 2004 (pp. 1367–1373). Geneva, Switzerland.

[4]Nasukawa, T. & Yi, J. (2003). Sentiment analysis: capturing favorability using natural language processing, In Proceedings of the 2nd international conference on Knowledge capture, October 23–25, 2003.(pp. 70–77).Florida, USA.

[5]Riloff, E., Wiebe, J., "Learning Extraction Patterns for Subjective Expressions", Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing(EMNLP), Japan, Sapporo, 2003