An Unsupervised Approach to Analyze Users Opinion on Products using Customer Reviews

Ravi kumar V[1] and Dr. K.Raghuveer[2]

[1] Research Scholar, Dept. of IS&E,NIE
Mysore, Karnataka, India

[2] Prof. & Head, Dept.of IS&E,NIE
Mysore, Karnataka, India

Abstract

Opinions are very much essential for any organization's development and success. It affects the quality of various products in a great way and will indicate how much the customer has satisfied with the product along with its strength and weakness. There are a number of ways of collecting opinions; it may be through a feedback register, or through forums and other online feedback survey facilities. The opinions are monitored and are used for improving the productivity of the product on which the opinion is received. After collecting opinions, it is important to analyze them. So we need a proper automated tool that can actually read through the comments and classify them according to the needs. The tool should be capable of going through a large set of comments in a short time and should be efficient enough to recognize almost any sentence we can throw at it.

Keywords— Document Classification, Opinion Identification, Product analysis, Semantic Orientation.

1. Introduction

The explosive increase in web communication has attracted technologies for automatically mining personal opinions from web documents such as product reviews and web blogs such technologies would benefit users who seek reviews on certain consumer product of interest [1]. On the other hand in order to improve the quality of the product and to keep the customers with more satisfaction, it has become essential to get the reviews from the customers about the products which they have purchased. This in turn enables the customers to review or to express opinion on products. Due to this we find lots of customer reviews about a product on the web, they can be used for analysis purpose to take necessary action. With all these it has become necessary to propose an enhanced approach for detection of web user's opinion on products.

2. Related Work

Many of the researcher's have proved that, it is possible to identify the opinion about a product by the customer using word at sentence level or at document level. Hence it is necessary to extract such words to detect opinion about a product.

"Extracting Aspect-Evaluation and Aspect-of Relations in Opinion Mining" [1], have focused on two important subtasks of opinion extraction: (a) extracting aspect-evaluation relations, and (b) extracting aspect-of relation. They adopted supervised learning method.

"Unsupervised Approaches for Detection of Web Users Opinion on Products" [2], it's an unsupervised approach for detection of web user's opinion. The approach identifies the user opinion on the basis of adjectives and classified as positive or negative using seed list.

"Extracting and Ranking Product Features in Opinion Documents" [3], focused on important task of mining features of an entity. The idea shows that identifying the features of a product is based on the opinion of user about that product. Using this features are ranked.

"Opinion extraction and summarization on the Web" [4], have focused on extracting the opinion about the different product features by the customer through the reviews obtained from the web. They use a supervised learning method to train the system with a specific pattern and from that they classify the opinion as either positive or negative.

3. Research Approach

Most of the time adjectives are the one which are used to express the opinion about any product or its features. It has

been observed that most of the times we find an adjective closure to Noun which may specify either the subject or the object, we are concentrating more on such patterns. Here we are proposing an unsupervised approach to detect the opinion of web users from the product review and to classify web user opinion into positive or negative. We use document based approach to classify the opinion by the people about a product as either positive or negative.

3.1 Document based Approach

The document based approach consists of five steps. In the first step the file containing the review will be modified to replace all words ends with "n't" with "not", for example "doesn't" will be replaced with "does not", so that it becomes easy to find the negation in the sentence. In the second step the modified file containing the review is tagged using Stanford tagger. In the third step the feature are extracted using tagged file by extracting the phrases based on the user defined pattern and are stored in a file. In the fourth step the extracted phrase is checked for the polarity. The polarity is either positive or negative based on whether the phrase detected is in the positive seed list or in the negative seed list. In the fifth step opinion is detected based on the score of the review, which will be calculated by considering all the features extracted. The flow chart for the above said steps is shown in Fig. 1.

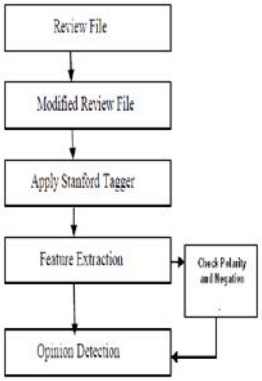

Fig.1: Flow chart for document based approach

Researchers have shown that either the adjectives or the adverbs are the one used to express their opinion about a product by a customer. Hence the algorithm extracts the phrases containing adjective or adverb. However Hatzivassiloglou, V *et al.,* have also developed an algorithm for predicting semantic orientation. Their algorithm performs well, but it is designed for isolated adjectives, rather than phrases containing adjectives or adverbs [5]. Therefore algorithm extracts the pattern containing adjective / a dverb and a co ntext word. The following patterns are extracted for extraction of phrases.

1) **Noun + Verb + Adverb + Adjectives:** The pattern consists of first noun and immediate verb followed by adverb if present and will stop when it finds an adjective. Here we are considering the adverb to find the negation of the adjective i.e. the presence of word "not" between noun and the adjective. For example "The battery is not good". Here "good" gives the opinion and "not" negates this opinion.
2) **Noun + Adverb + Adjectives:** The pattern consists of first noun and followed by an adverb if present and will stop when it finds an adjective. Here we are considering the adverb to find the negation of the adjective i.e. the presence of word "not" between noun and the adjective. For example "The battery not good".
3) **Adverb + Adjectives + Noun:** In this case we first identify the noun and then we will move towards back to find adjectives adjacent to this noun and then we will search for adverb representing the word "not" with a window size of 5, to the left of noun. For example "Nice cell phone".
4) **Prep + Verb + Adverb + Adjectives**: First word will be the preposition, followed by immediate verb and continues still you find an adjective, however we stop when we find a noun. Similar to the previous pattern we consider the adverb between the first word and the last word, if present.
5) **Adverb + Adjective**: This pattern is considered to identify any of the missing adjectives from all the above mentioned patterns. In this case we identify the adverb and immediate adjective.
In addition to this we have considered the adjective which comes next to the one which we identified, in the pattern 1, 2, 3 and 4.

The algorithm extracts the above mentioned patterns and they will be stored in a file. During this process care has been taken so that there should not be any repetition of the adjectives. Similarly we are checking for the presence of negation within the pattern.

Check for polarity- To find the sentiment of an opinion word fetched in the previous step, it is necessary to have lexicon. In order to build this lexicon we have used following Lexical Resources for Opinion Words:

i) Subjectivity Lexicon: The list Subjectivity clue (the subjectivity lexicon) is part of OpinionFinder and was supported in part by NSF Grants, and it can be downloaded at http://www.cs.pitt.edu/mpqa/lexiconrelease. From this we have considered 2304 pos itive words, 4143 number of negative words and the remaining neutral, words we have not taken into consideration [6].

ii) General Inquirer(GI) : It is web based system which gives a list of words in different category with small description about each word. For example for the word 'difficult' the GI gives "difficult Neg Modif Vice EVAL Ngtv | adjective: Not readily done, demanding, hard ". This provides information like different senses of each word along with the polarity. Some of the word will have different senses, in that case it provides frequency of usage for each senses and higher frequency will be considered. We have considered 77 positive and 45 n egative words labeled with positive or negative words respectively and along with they are labeled by 'Modif', because most of the time opinion word will be a modifier of some feature.

iii) SentiWordNet : It is a lexical resource explicitly devised for supporting sentiment classification and opinion mining application[7]. This gives three scores: positive score, negative score and object score, for a given word based on the specified POS. For some of the word it gives more than one above said scores, but we have chosen the first one, because these values are arranged based on the frequency i.e. first one will have more frequency than the other.

Along with the positive and negative words obtained from Subjectivity Lexicon and General Inquire to the seed list, we have identified some of the words considering the domain even though they are available in the GI and Subjectivity Lexicon, from the customer opinion point of view they have opposite polarity and are important from the point of opinion, such words are added to the seed lists and also some of the words neither they are in the GI and nor in the Subjectivity Lexicon, but they are specific to the domain, we felt by adding these words to the seed lists, it becomes more strong, hence we have added very few number of p ositive and ne gative words in to the respective seed lists.

During this process we found that some of the words (adjectives) will not give any meaning in finding the opinion, such words have been listed under domain constraint list; these are used in the process of finding polarity of an opinion word.
We use Positive seed list, Negative seed list and domain constraint list to find the polarity of the opinion word obtained from the previous step. The extracted features from the previous step are stored in a file and it is used as input. The Fig. 2 shows the flowchart to find whether the adjective is positive or negative or neutral.

In the first part of the flow chart, the adjective (opinion word) will be read from the input file, and it will be searched under the above mentioned list and also it will be checked for neutral polarity under Subjectivity Lexicon. Based on its presence it returns either 0, or 1 or -1. If it is not present in any of these, then in the second part we use SentiWordNet, it will returns three numerical scores, one for Positive, one for Objective and another for Negative. If the Positive score is greater than 0 and it is greater than Negative score, then it will be searched under Positive seed list if it is present, it will return 1 or else it will return 0. Similarly if the Negative score is greater than 0 and it is greater than Positive score, then it will be searched under Negative seed list if it is present, it will return -1 or else it will return 0. If the positive score and Negative score both are equal to 0, then we use WordNet, to get the first 2 Synonyms of that word, if it returns null, then the word will be discarded, otherwise if at least one of the synonym matches with the Positive or Negative seed list, the respective value will be return or else it will return 0.

If the adjective is present in the positive seed list, the algorithm returns 1 and positive score will be incremented by 1. Similarly algorithm returns -1 if the adjective is present in the negative seed list and negative score will be incremented by 1. During this process we check for the presence of negation i.e. "not, before the adjective in each of the pattern extracted in the previous step. In patterns 1, 2 and 4 the word "not" will be checked between first word and fourth word, but in case of pattern 3 and 5 a window size of five has been taken to the left of adjective and checked for the presence of word "not". If "not" found then the polarity is negated by multiplying it by -1.

Opinion Detection- Once the polarity has been known, the next step is to detect whether the document gives either a positive opinion or a negative opinion. In order to detect this score of a document is obtained by adding the positive score and negative score for each of the feature extracted. We have taken 0 as threshold T. If the score is greater than the threshold value we consider the opinion as positive and if the score is less than the threshold value we consider the opinion as negative.

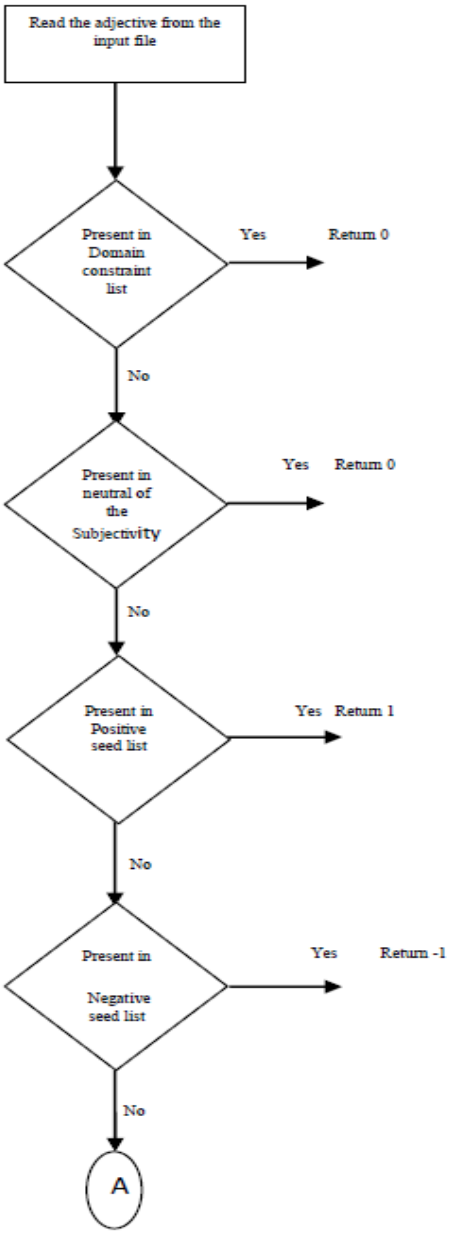

Fig.2: Flow chart to check the polarity of an adjective, Contd.

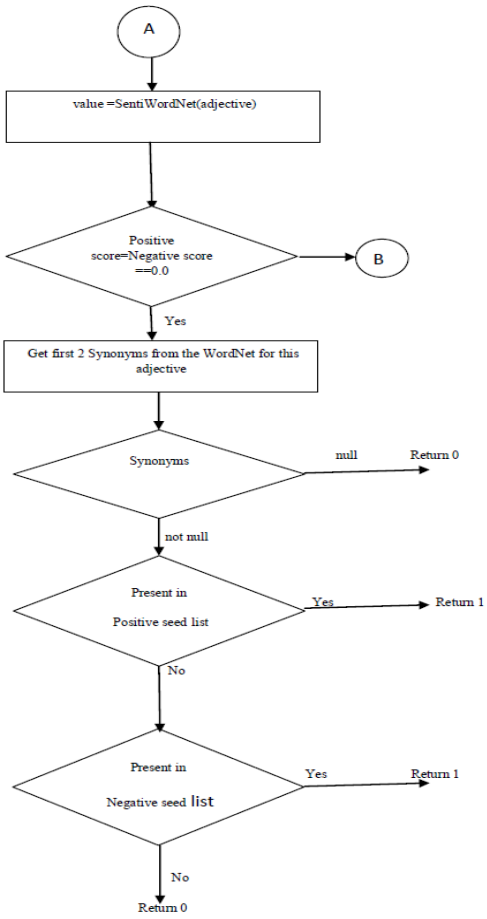

Fig.2: Flow chart to check the polarity of an adjective, Contd.

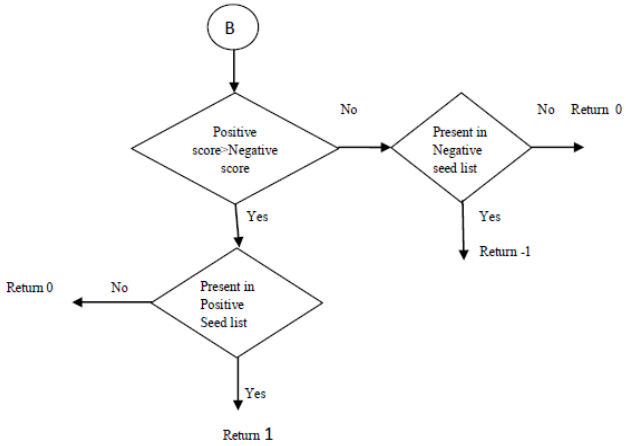

Fig.2: Flow chart to check the polarity of an adjective

4. Experiments and Results

A. Data Sets

We have used two different data sets to evaluate our approach. The data sets consist of customer reviews (more than 450 reviews) of two electronic products from web sites like Amazon.com, Cnet.com:2 digital cameras and 1 DVD Player.

We have conducted various experiments with our approach. The details of the experiment and corresponding results are discussed below.

Experiment 1: Feature extraction using Document based Approach.

Each customer review obtained from various web sites was stored in a separate file. In this experiment each file will be given as input at a time, if the score is greater than the threshold value, the opinion is considered as positive, otherwise if the score is less than the threshold value, the opinion is considered as negative.

Experiment 2: Classification of Multiple Reviews on Data Set 1.

In this experiment the proposed approach is tested against multiple reviews of product on data set 1 and the detection of opinion in terms of percentage is calculated.

$$\text{The Accuracy} = \frac{\text{The total number of documents has identified correctly as positive + The total number of documents has identified correctly as negative}}{\text{The total number of documents to be classify as either positive or negative.}}$$

Table 1 shows the product wise reviews considered for classification and they are classified as positive or negative using Document based Approach using data set 1.

Table 1: Classification of Multiple Reviews using Document based Approach on Data set 1

Sl. No	Product	Number of Customer Reviews	Positive Accuracy (%)	Negative Accuracy (%)
1	Canon Digital Camera	35	100	48.57
2	Nikon CoolPix Digital Camera	30	100	50.00
3	Samsung DVD	38	97.36	60.52

The graph shown in the Fig. 3 indicates the positive and negative accuracy for multiple documents on different products using Document based approach on Data set 1. X-axis shows the different products and Y-axis shows the accuracy in %.

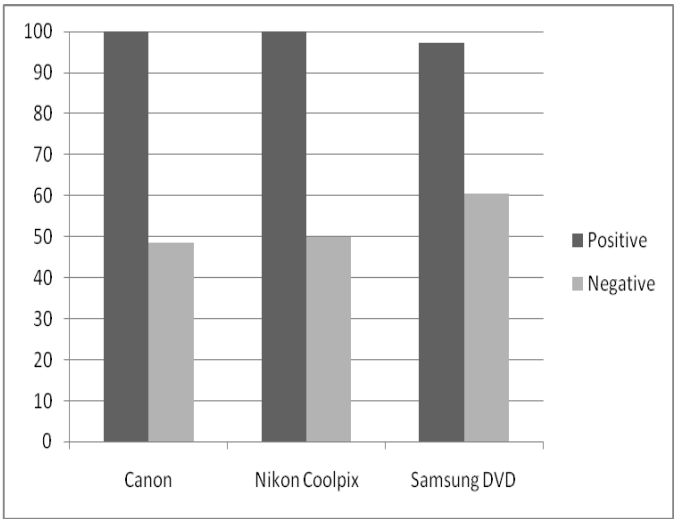

Fig.3: Accuracy of approach on different products for Data set1

Experiment 3: Classification of Multiple Reviews on Data Set 2.

Table 2 shows the product wise reviews considered for classification and they are classified as positive or negative using Document based Approach using data set 2.

Table 2: Classification of Multiple Reviews using Document based Approach on Data set 2

Sl.No	Product	Number of Customer Reviews	Positive Accuracy (%)	Negative Accuracy (%)
1	Canon Digital Camera	35	100.00	40.00
2	Nikon CoolPix Digital Camera	30	96.66	56.66
3	Samsung DVD	38	94.53	44.77

The graph shown in the Fig. 4 indicates the positive and negative accuracy for multiple documents on different products using Document based approach on Data set 2. X-axis shows the different products and Y-axis shows the accuracy in %.

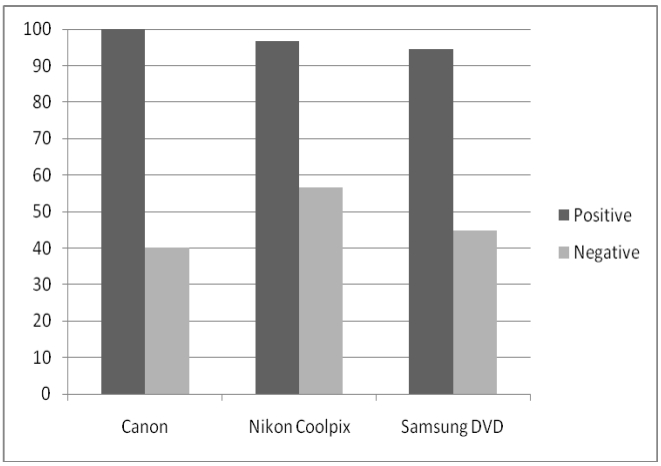

Fig.4: Accuracy of approach on different products for Data set2

5. Conclusions and Scope for Future

The research results have shown a promising opinion mining approach using customer review on different products, which is important from the point of individuals and organization. However, this approach still has the opportunity for great deal of improvement. There are three main courses of action for future work to reap the benefit of this approach: Analyze the current result and extending this approach with the help of some dependency parser. Generation of more accurate seed lists automatically. Use of other part-of-speech (POS) along with adjectives to improve the accuracy.

REFERENCES

[1] Nozomi Kobayashi, Kentaro Inui, and Yuji Matsumoto. "Extracting Aspect-Evaluation and Aspect-of Relations in Opinion Mining". Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational. Natural Language Learning, pp. 1065–1074, Prague, June 2007.

[2] Anil Kumar .K.M, Anil Kumar .P, and Suresha."Unsupervised Approaches for Detection of Web Users Opinion on Products". Journal of Computing, Volume 2, Issue 8, August 2010, ISSN 2151-9167.

[3] Lei Zhang, Bing Liu, Suk Hwan Lim, and Eamonn O'Brien-Strain. "Extracting and Ranking Product Features in Opinion D ocuments". Colling 2010: Poster Volume, pages 1462-1470, Beejing, August 2010.

[4] Minqing Hu, and Bing Liu. "Opinion extraction and summarization on t he Web". American Association for Artificial Intelligence.

[5] Hatzivassiloglou, V. and K. McKeown 1997. " Predicting the Semantic Orientation of Adjectives". Proceedings of the 35th ACL conference, 174–181.

[6] Anil Kumar K. M and Suresha. "Identifying Subjective Phrases from Opinionated Texts using Sentiment Product Lexicon". P ublished in International Journal of Advanced Engineering & Application, Jan. 2010

[7] Hanjie Shu "Opinion Mining for Song Lyrics" Norwegian University of Science and Technology Department of Computer and Information Science

[8] Ellen Riloff. "Automatically Constructing a Dictionary for Information Extraction Tasks". Proceedings of the Eleventh National Conference on Artificial Intelligence, 1993, AAAI Press / MIT Press, pages 811–816.

[9] Nozomi Kobayashi, Tyu Iida, Kentaro Inui, and Yuji Matsumoto. "Opinion Extraction Using a Learning-Based Anaphora Resolution Technique". The 7th Message Understanding Conference (1998).

[10] Nobuaki Hiroshima, Setsuo Yamada, Osamu Furuse and Ryoji Kataoka. **"**Searching for Sentences Expressing Opinions". Workshop on Sentiment and Subjectivity in Text, pages 39– 46, Sydney, July 2006.

[11] Mathias Bank and Michael Mattes. "Automatic User Comment Detection in Flat Internet Fora". 20th International Workshop on D atabase and Expert Systems Application .IEEE Computer Society, Pg, 373-377

[12] Michael Thelen and Ellen Rilo. "A Bootstrapping Method for Learning Semantic Lexicons using Extraction Pattern Contexts". Proceedings of the 2002 C onference on Empirical M ethods in Natural Language Processing (EMNLP 2002)

[13] Minqing Hu and Bing Liu. "Opinion Feature Extraction Using Class Sequential Rules". American Association for Artificial Intelligence

[14] Peter D. "Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews". Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 417-424.

[15] Raymond J. Mooney, Razvan Bunescu, "Mining Knowledge from Text using Information Extraction", ACM SIGKDD Explorations, 7: pp 3-10

[16] Soo-Min Kim,Eduard Hovy, "Determining the Sentiment of Opinions " Proceedings of the COLING conference, Geneva, 2004

[17] Wai Lam and Kon-Fan Low, "Automatic Document Classification Based on Probabilistic Reasoning: Model and Performance Analysis". 1997 IEEE.

[18] Yi, J., Nasukawa, T., Bunesu, R., Niblack, W. "Sentiment Analyzer: Extracting sentiments about a given topic using natural language processing techniques". ICDM 2003

[19] Yi, H., and Hatzivassiloglou, V. "Towards answering opinion questions: Separating facts from opinion and identifying pol arity of opinion sentences". EMNPL 03 2003.