International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Shopping Site Recommendation Using Lexicon-based Approach of Sentiment Analysis

Nidutt Nilay Bhuptani[1], Smitha B[2], H M Pooja[3], Shashank Shekar[4]

[1,2,3,4] Department of Computer Science and Engineering, PESIT-BSC Bangalore, India
[1]niduttnb@gmail.com, [2]bsmithagowda@gmail.com, [3]poojahm3@gmail.com, [4]shekar.shashank1122@gmail.com

Abstract: ***In recent years Sentiment analysis of how customers describe or talk about a given product or brand has gain much attention as the retail environment becomes more competitive. Sentiment analysis identifies and categorizes the opinions expressed in the texts. It is used in order to determine writers attitude towards the topic, product etc. In current analysis of social media comments are used as a basis to recommend best shopping site. In proposed system reviews of different E-shopping websites are collected and analyzing which shopping site is the best based on reviews. We are using Sentence level sentiment analysis to classify the sentiment expressed in each sentence. It determines whether the sentence expresses positive or negative opinions. We aim at creating lexica in which opinion expressions are annotated according to their polarity, using the dictionary. We have created dictionary manually. We are calculating the orientation for an opinion based on the positive and negative words that we have in the dictionary.***

Keywords: Sentiment analysis, lexica, E-shopping websites, reviews analysis

1. Introduction

Online shopping has become the integral part of retailing industry. The main attraction for the customers in online shopping is convenience. Tons of people buy goods online every day. The abundant amount of user data is generated in E-shopping websites or application in the form of reviews which can be used to recommend best shopping site by analyzing user reviews. Product reviews are powerful and also provide clear picture of the product. Websites like Jabong, Amazon, Snapdeal, Flipkart, Ebay etc are popular sites where millions of users exchange their views, opinions and making it a valuable platform for analyzing and tracking opinion and sentiments.

In the field of text mining sentiment analysis is one of the ongoing fields of research. Sentiment analysis is the field of study that analyzes people's sentiments, opinions, emotions and evaluations towards entities such as products, organizations, services, issues, individuals, topics, events, and their attributes [1]. Sentiment analysis helps to understand user's attitude, emotions and opinion towards the entities. There are three classification levels in sentiment analysis, aspect level, sentence level, and document level sentiment analysis. A sentence level sentiment analysis is used in our work, it classifies expressed sentiment in each sentence to determine whether the sentence express positive or negative opinions.

Application of a lexical analysis is one of the two main approaches to sentimental analysis. There are three classification techniques in sentimental analysis, lexical based approach, machine learning approach and hybrid approach [2]. The machine learning approach uses the machine learning algorithms. The hybrid approach combines both machine learning and lexicon based approach. The lexicon based approach involves calculating the sentiment from the semantic orientation of word or phrases that occur in a text [3]. The lexicon based approach is divided in to dictionary based approach and corpus based approach. Dictionary based approach of lexicon analysis is used in our work. With this approach a dictionary of positive and negative words is required which can be created using manual [4] or an automatic [5] approach. In many sentiment classification tasks opinion words are employed. Positive words are used to tell some desired states and while negative words are used to tell some undesired states.

In this paper, we have used five dataset of five different E-shopping websites i.e. Amazon, Snapdeal, Flipkart, Ebay and Jabong. These contain reviews related to the mobile phones and then classify them into positive, negative and neutral. After analysis of these reviews we suggest best shopping site to the user for a particular product.

The rest of the paper is organized as follows: Section II discusses the related works done previously by different researchers. Section III presents an implementation of proposed system. Section IV discusses about data analysis and results. Section V discusses about the results obtained. Finally, this paper is concluded by suggesting some possible future work.

2. Related Work

A lot of work has been done in the area of sentimental analysis based on Lexical approach. This section discusses about few of the works related to this domain.

A novel dictionary-based algorithm that uses lexicon-based approach for opinion mining to calculate the sentiments through a text by building a dictionary of sentiment words with three degrees of comparison viz. positive, comparative and superlative is proposed in [6].

Focusing on handling polarity shift problem, in paper [7] they have proposed a model called dual sentiment analysis (DSA), to classify the reviews by considering two sides of one review. This address the problem of polarity shift in sentiment classification.

The sentiments of reviews from five different e-shopping sites, collected from online source are analyzed in [8]. Preprocessing techniques are used to remove unwanted things from reviews. "Sentiwordnet dictionary" is used for finding score of each word in review. Then sentiments are classified as positive, negative and neutral. It is observed that the quality of detected sentiments is greatly affected by the pre-processing of the data.

To detect sarcasm on Twitter they have proposed a pattern-based approach in [9] which it classify each tweet depending on whether it is sarcastic or not from the set of given tweets. The features are extracted in such a way that it covers different types of sarcasm by making use of different components of the tweet.

To predict the sentiment behind a status post of Facebook which is in the nature of cross language domain, unstructured dataset and noisy in [10] they have used lexicon based dictionary approach. Since traditional opinion mining is not efficient enough to find out sentiments from status post of social media's like Facebook. A group in Facebook called "foodbank" is choosed to analyze sentiment of post in order to find out their market values.

Focusing on problem of predicting rating based on the comments of internet users, in [11] they have proposed a classifier which is based on Vector Space Model and information retrieval in order to solve problem and they also investigated on effect of integrating sentiment analysis model into the classifier.

Recently, the interest of sentiment analysis in social media has increased. Social media is the primary source of big data. Social media data are noisy and unstructured. The combination of Lexicon based dictionary approach and support vector machine have been applied for assessing the performance of television program in Twitter [12].

Sentiment analysis has many applications on internet. Analyzing movie review is one of them. Online reviews contain both subjective and objective sentences. Objective information does not contain sentiment or opinion therefore only subjective information is useful in such cases [13]. Naïve bayes classifier is applied to extract subjective sentences. Naïve Bayes classifier gives more accurate result than SentiWordNet.

Sentiment analysis can be done in machine learning approach, lexicon based approach and hybrid approach. The consequences and viewpoints of dictionary based approach are discussed in [14].

Lexical-based approach is used extract sentiment from the text. Semantic Orientation CALculator (SO-CAL) is applied for polarity classification [15]. The SO-CAL is consistent in many domains. Performance metric is better compared to traditional methods.

3. System Architecture

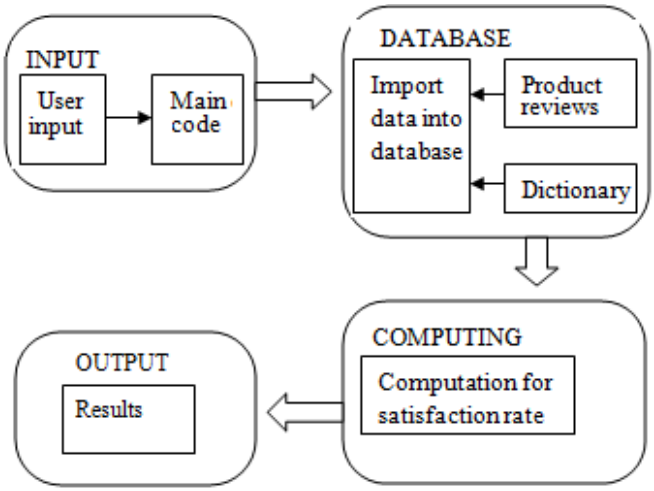

Figure 1: System architecture

3.1 Input

Input is given as product name entered by the user for which he needs to know best shopping site for a particular product.

3.2 Database

Database mainly consists of data from two components; one is product reviews which contain reviews of product from different websites in order to analyze sentiment behind reviews and another is dictionary of positive and negative words as we are using dictionary based approach.

3.3 Computing

Using data in database, reviews from different sites are analyzed. After analyzing satisfaction rate is computed for each website.

3.4 Results

Bar graph indicating satisfaction rate of each website is displayed as a result.

4. Implementation

The implementation stages of the work is a typical example of Lexical based approach wherein two different dictionaries i.e. positive and negative are formed and the data set is looked for the words in the dictionary. The best part about the method adopted by us is with least computation time, it gives the best results. Although the average computation time is a function of the amount of data set, on an average we obtained results for any product in less than a second. Figure 2 depicts the flow of the work.

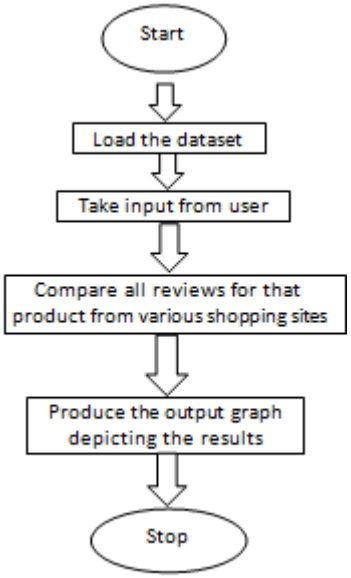

Figure 2: Flow Chart

1) The data set is abstracted using appropriate tools of data mining. The extracted data is stored in a '.CSV' file. In the data set, reviews should be stored in the same row for each product. This smoothens the process during data analysis of it.
2) Next, the dictionary of positive and negative words is to be formed. It is stored in '.TXT' format .The dictionary is the keystone in the accuracy. More far-flung and accurate the dictionary, the more efficient and accurate the result would be retrieved.
3) The user then enters the product name which needs to be analyzed. This product is then matched against the products in the data set. After a match is found the corresponding reviews are sequentially analyzed.
4) During the analysis, the words in the reviews are checked against the words predefined in the dictionary i.e. a simple string matching operation is carried out cumulatively. The counter for both positive and negative words is kept, which helps for the satisfaction level calculation.
5) Now based on the counter values, the satisfaction percentage is computed.
6) The computed result is then presented in the form of a webpage, where the user inputs the product name in a search bar and using jQuery the user is redirected to the page of the product the user. Non existence of the product entered by the user would display as 'No such product found'.
7) The result should preferably be in the form of a bar/column graph which clearly demarcates the result so that even a layman can understand the result.

The implementation of such system helps the customer to save their searching time for purchasing products through online shopping sites. Prior to this, the customers visits all the shopping websites, compare their reviews and then choose a product. Our implementation gives the user a perfect idea of the best rated website in the form of a clear bar graph.

Let's assume the counter values for positive statements/phrases/words is x and for negative is y. The following is the equation we derived where S is satisfaction rate: $S = (x / x + y) * 100$

5. Results and Analysis

The purpose of this project was to help the user to choose the best shopping site and evaluate the performance of our lexicon based SA algorithm.

We have summarized the result in in Table A (Precision Calculation) and on observing the table it pretty evident that performance has been satisfactory.

Table 1: Precision Calculation

Polarity	Positive	Negative	Neutral	Precision
Positive	410	13	20	92.5%
Negative	7	18	4	62.06%
Neutral	11	5	16	50%

Now for each product, based on the reviews in the dataset the overall result is displayed in the form of bar graph as shown in Figure.3.

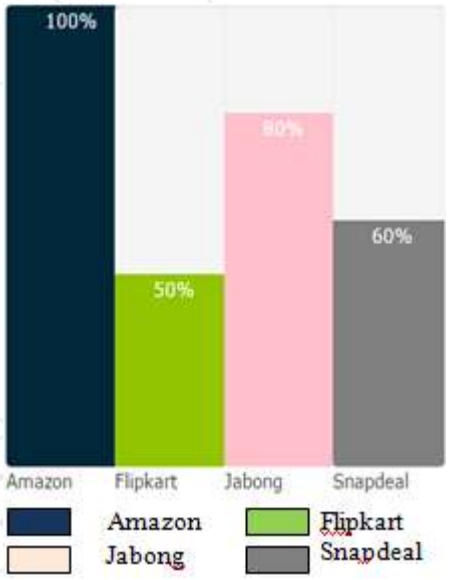

Figure 3: Output Graph

The result clearly helps the user to decide the best shopping site to choose product from. After finding out the result for around 80-100 products, we found out the accuracy percentage to be around 85-90%.

The algorithm failed to deliver accurate results in certain cases. These were the following cases-

1) The reviews don't have correct spellings. In these cases since there are so many wrong spellings for a words, it is almost impossible to have them all in the dictionary.
2) The reviews are sarcastic which is quite common these days. The exact meaning was not able to be interpreted.

A typical example of result obtainer for one particular product is shown above. It is in the form of a bar graph. Over here, the following observations are notes-

1) The users are most satisfied with Amazon. The users are also satisfied with Jabong.
2) However, customers were not happy with Flipkart.
3) So as a result anyone will opt for Amazon.

Note that if the satisfaction percentage obtained by this method is 50%, it can imply two things-

1) The review of the user is ambiguous, i.e. the review is neutral OR
2) No sentiment of user was depicted by our system. Therefore by default in all cases, it is set to 50% by us. This mainly happens if the spellings used by users are completely wrong and so the key words couldn't match with our dictionary. This was the only case where our system wasn't able to analyze the exact sentiment of the user in the review.
3) Summarizing it, for most of the cases, we were able to understand the sentiments of the review.

6. Conclusion

In this paper, we have shown how to predict sentiment behind customer reviews of products in E-commerce websites using lexicon based approach. With positive, negative and neutral dictionary reviews are classified accordingly to predict the customer satisfaction rate. This work gives high accuracy of around 85-90%. It saves the customer time by recommending best shopping site after analyzing reviews from websites.

Our work, which is restricted to phones only, can be extended in future for any other products there on the internet. It can be further extended by keeping in mind the various other parameters like cost, delivery time etc. Sentimental analysis is still quite unexplored. It can be used as an important tool for lot of purpose like analyze movie reviews, on movie booking sites like Bookmyshow, analyze the outcome of elections based on how many positive reviews are there on twitter about that candidate, Gaming companies can use it to see how the gaming community is reacting to the latest launched game and see what bugs needs to be fixed or which extra features are required to be added, In case the government makes a new plan, sentiment analysis on twitter data can be carried out to see the people are reacting to it, Social media companies can see if they their newly launched feature went well with the public or not etc.

7. Acknowledgment

We would like to thank our peers Mohammed Yaseen and Shaikh Muzzamil who are active participants in implementation process and we also thank our Prof. Saraswathi Punagin whose moral support and encouragement lead to the success of the paper.

References

[1] B. Pang, L. Lillian, "Opinion mining and sentiment analysis," Foundation and trends in information retrieval, vol. 2, 2008.
[2] W. Medhat, A. Hassan, H. Korashy, "Sentiment analysis algorithms and applications: survey," Ain shams engineering journal, 2014.
[3] M. Taboada, J Brooke, M Tofiloski, K. Voll, M. Stede, "Lexicon-Based methods for sentiment analysis," Association for Computational Linguistics, vol. 37, pp. 267-307, 2011.
[4] RM. Tong, "An operational system for detecting and tracking opinions in on-line discussions," Working notes of the SIGIR workshop on operational text classification, 2001.
[5] P. Turney, M. Littman, "Measuring praise and criticism: Inference of semantic orientation from association," ACM Transaction on information systems, vol. 21, pp. 315-346, 2003.
[6] Tiara ,S. Mira Kania, E. Veronikha, "Sentiment Analysis on Twitter Using the Combination of Lexicon-Based and Support Vector Machine for Assessing the Performance of a Television Program", 3rd International Conference on Information and Communication Technology (ICoICT), 2015, pp. 386-390.
[7] M. Santanu, G. Sumit, "A Novel dictionary-based classification algorithm for opinion mining," Second international conference on research in computational intelligence and communication networks, 2016.
[8] Rui Xia, Feng Xu, Chengqing Zong, Qianmu Li, Yong Qi, Tao Li, "Dual sentiment analysis: considering two sides of one review," IEEE Transactions on Knowledge and Data Engineering, vol. 27, pp. 2120-2133, 2015.
[9] U. Ravi, "Sentiment Analysis of Reviews for E-Shopping Websites," International Journal of Engineering and Computer Science, vol. 6, pp. 19965-19968, 2017.
[10] Mondher Bouazizi, Tomoaki Ohtsuki, "A Pattern-Based Approach for Sarcasm Detection on Twitter," IEEE Access, vol. 4, pp. 5477-5488, 2016.
[11] A. Sanjida, T. Muhammad, "Sentiment Analysis On Facebook Group Using Lexicon Based Approach," electrical engineering and information communication technology, 2016.
[12] Eissa M. Alshari, Azreen Azman, Norwati Mustapha, Shyamala C Doraisamy, Mostafa Alksher, "Prediction of Rating from Comments based on Information Retrieval and Sentiment Analysis," Third International Conference on Information Retrieval and Knowledge Management, 2016.
[13] B. Purtata, K. Shilpa, "Sentiment Analysis of Movie Reviews Using Lexicon Approach", IEEE International Conference on Computational Intelligence and Computing Research, 2015.
[14] C. Fatehjeet Kaur, B. Rekha, "Sentiment Analyzing by Dictionary based Approach", International Journal of Computer Applications (0975 – 8887, Vol 152 – No.5, October 2016, pp. 32-34.
[15] T. Maite, B. Julian, T. Milan, V. Kimberly, S. Manfred, "Lexicon-Based Methods for Sentiment Analysis", Association for Computational Linguistics, Vol 37-No 2, 2011, pp. 267-307.