International Journal of Computer Sciences and Engineering Open Access
Survey Paper Volume-4, Special Issue-7, Dec 2016 ISSN: 2347-2693

Customer Opinion from various E-commerce sitesusing Data Mining techniques: A Survey

PinkySaikiaDutta[*1], AvishekSaha[*2], AmarendraSharma[*3], AbhishekSarmah[*4], JugalKishorTalukdar[*5]

Department of CSE, GirijanandaChowdhury Institute of Management & Technology
Hathkhowapara, Azara, Guwahati-17
[1]pinky_sdutta@rediffmail.com
[2]avishekk2013@gmail.com

Available online at: www.ijcseonline.org

***Abstract*— With the availability of huge number of products, it became quite difficult for a customer to judge the quality about the product. Publicly available opinions are very important for decision making process. With the increasing number of reviews and comments for a particular product it became difficult to get an optimized opinion for that product. In this paper, we will study various reviews and the quality of the reviews from a huge number of positive and negative opinions on the product. All the reviews will be analysed on the basis of sentiment and will give a final opinion on the product. It will help every buyer to take a quick decision and gain a precise opinion for the products.**

***Keywords*—Optimized opinion; Quality; Sentiment.**

I. INTRODUCTION

Now a day, viewing opinions before buying product has become an increasingly popular way to know wellabout products. Posting reviews has become a popular way to express opinion about products. Reviews on particular product consistently increases. Analysing from a large volume of opinion and reviews becomes quite difficult for a particular user to take a decision for buying products. For example, Amazon.com archives a total of more than 480 million products. Bing Shopping has indexed more than five million products. Shopper.com records more than five million products from over 3,000 merchants [1]. It becomes probably impossible for any individual user to go through all reviews.Moreover, in various website user share opinion on various product. Considering amazon.com and flipkart.com there are various types of products available and for per products thousands of opinions are available. So, it's a tough task to go through all these opinions for getting a real and optimized result.

Opinions or review on a product can be of any type like good review, bad review, true review or fake review. So, retrieving a real or optimized review has become really a challenging task. Maximum percent of reviews are in the form of text which are basically available in various websites. So, extracting a real and optimized opinion involves various text mining techniques. Many current sites use star rating system to make quality judgment of the product and to get customers view on that product [2]. It is efficient to a certain extent. Users get an overview of a product but cannot get an exact view for what the product is given such rating. What features of the product is of good quality and for what it is preferred as such? What features of it is bad. For example, considering a mobile, which is of good quality as a whole but it may just have a small fault. Now a customer neglecting that fault can give a fine rating e.g., 5/5, but it is not given the proper view. Further, if he gives a bit low rating e.g., 4/5, but it may happen that the mobile comparing with other mobiles of the same rating can give better user performance. So, it cannot be properly judged by the rating review of the product. So, it is better moving to text reviews. In text based reviews, customers give their views or opinions in the form of text where a buyer can get the exact views of the product. For this we have to move up to sentiment analysis.

II. RELATED WORK

In this portion, we will relate our study work with the existing works.

Regarding opinion analysis, the first step is to collect various opinion or review from various websites like amazon and flipkart. For this, extraction of data is done through crawling. Crawler is required to crawl the various sites over the internet, basically the sites which we target. The crawler is set for extracting various product information data including html tags, comments, reviews, etc. All this informationare raw data, semi structured and unstructured data. Raw data are preprocessed.

Minqing Hu and Bing Liu [3] clarify that sentiment analysis work is more fruitful after preprocessing. They discussed about POS tagging technique. POS tagging means word-category disambiguation. Using this technique, they label various words under a common label or category. They used NLProcessor linguistic parser to parse every review and producing part of speech tag for each word like nouns, verbs and adjectives. Each review along with the other POS tagged information is saved in the database. These sentences contain nouns and noun phrases. The remaining components are product features.

DrS Vijayarani, Ms J Ilamathi,Ms. Nithya discussed [4] shows the removal of suffixes from different words to

have accurately matching stems. They used two themes during stemming: Words not having the same meaning to be kept separate and the different morphological forms of texts should be mapped into a same stem. This process reduces the number of words and memory space.

Theresa Wilson, JanyceWiebe, Paul Hoffmann [5] states that Sentiment analysis is the task of recognizing positive and negative opinions, emotions, and evaluations. This classification is done to classify the given text into three levels - document level, sentence level and aspect level. Document level sentiment analysis classifies an opinion as expressing a positive or negative opinion. Sentence level classify sentiments in each sentence. Mining of reviews require sentence level sentiment analysis. Document level and sentence level classification does not provide necessary detail needed opinions on all features of the products. So, aspect level sentiment analysis is needed to obtain these details [6].

Lina L. Dhande and Dr. Prof. Girish K. Patnaik [7] used the Naive Bayes Neural Network Classifier to improve the accuracy and performance of Sentiment Classification in real world dataset. They referred Naive Bayes as the simple probabilistic classifier based on applying Bayes theorem which works well on text classification. it shows that all attributes are independent given the value of class variable i.e., each feature is conditionally independent to the features given the class. The Bayes rule is given as

$$P(c|t) = P(c)* P(t/c) / P(t)$$

Where, c is a specific class and t is the text to be classify. They mentioned the advantages as simple, fast and high accuracy. Secondly they took into account Neural Network Classifier. The neural network with network structure shows the dependency among various input variables. They mentioned the limitations of the Naive Bayes Classifier on NaiveBayes Neural Classifier.

Lexicon based approach is used to extract sentiments from text.ChetanKaushik and Atul Mishra[8] classified the extracted texts into positive, negative and neutral opinion with the help of a dictionary having sentiment bearing words along with their polarities. In this paper, it showed the limitations of Machine Learning Technique based on the basis of performance in time. In Sentiment Lexicon Technique component, the words are compared with the dictionary of specific domain which contains all forms of words i.e. every word is stored along with its various verb forms e.g. applause, applauding, applauded, applauds and contains the strength of the polarity of every word. Some word depicts stronger emotions than others. Negation and blind negation words are are also used for identifying the sentiments in the sentence, as their presence can reverse the polarity of the sentence.Thedictionary also contains several different adjectives, nouns, negation words, emoticons etc

SVM is a supervised machine learning algorithm which is mostly used for classification. Ms. GaurangiPatil, Ms. VarshaGalande, Mr. VedantKekan, Ms. KalpanaDang [9] used SVM technique to compare the different reviews which mean the same. There are also certain reviews which become difficult to decide whether this is a positive or a negative review.Product review may show negative impact but if it is seen logically it is a positive one. These are some approaches:TP (True Positive), FP (False Positive), TN (True negative) and FN (False negative) for categorization of the reviews.

Haiyung Sui ChristoperKhoo, Syin Chan [10] shows that the use of unigram approach which shows the most accuracy in traditional topic text classification.

In study from Vijay Murari T, Sunil, Suchitra [11] the Maximum Entropy classifier is a probabilistic classifier which belongs to the class of exponential models. The Maximum Entropy is based on the principle of maximum entropy and from all the models that fit our training data, selects the one which has the largest entropy. Wide range of text classification can be done using maximum entropy such as sentiment analysis. For classification, maximum entropy is used to framework for integrating information from many heterogeneous information. There are varieties of maximum entropy classification problem in natural language processing, such as sentence boundary detection, information extraction, and part of-speech tagging. In classification, Maximum entropy works better than naïve Bayes classifier. Maximum Entropy is used to extract product review from e-commerce site. Maximum entropy outperforms Naive Bayes at standard text classification.

S. ChandraKala and C. Sindhu [12] describes that Maximum Entropy makes no independent assumptions about the occurrence of features and when conditional independence assumptions are not met it might perform better.

Anascollomb, CrinaCostea, Damien Joyeux, Omar Hasan, Lionel Brunie[13] expressed Unsupervised Learning has no target attribute.

So, it does not have any training set. It explores or processes all the data and finds some intrinsic structure in them. Unsupervised learning mainly performs clustering a classification. Applying unsupervised machine learning data are organized into class such that their is

a) High intra class similarity

b) Low intra class similarity

Unsupervised learning on row (not preprocessed data) data may because some error in accuracy achievement. So,already preprocessed data are taken for unsupervised learning. So, it gives some high accuracy. Previously in preprocessing task the noun, verb and adjective are POS tagged and other data are taken as unlikely. Along with that stop word removal and stemming is performed. So, remaining text collection are organized according to their content similarity. To produce a topic hierarchy, in fact clustering is one of the most utilized techniques for unsupervised learning. Because it provides the features of

- Scalability,
- ability to work on unknown type of data set,
- don't need to focus on input parameter,
- able to deal with noise and anomaly
- Interpretability and usability.

Based on clustering label and the classification label the polarity can calculate which gives better result on sentiment classification like.

MaitebTaboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, Manfred Stede[14] described that most of the majority of statistical text classification research builds SVM (Support Vector Machine) classifier, trained on a particular data set using features such as unigrams or bigrams, and with or without part-of-speech labels, although the most successful features seem to be basic unigrams. Classifiers built using supervised methods reach quite a high accuracy in detecting the polarity of a text. However, although such classifiers perform very well in the domain that they are trained on, their performance drops precipitously (almost to chance) when the same classifier is used in a different domain.

III. EXPERIMENTAL METHOD

We approach above methods for better accuracy for measuring sentiment. In the fig-1 we show the work framework for obtaining optimised result from product review.

Data are collected using various crawler's API's and are stored in a database. After collection, we are going to pre-process the stored data. This includes removal of all the noisy and unnecessary data and POS tagged all the sentences. Here we are going to acquire data that are required for our analysis.After that we are going to apply sentiment classification algorithm to obtain the exact opinion for the review. After that we are going to apply naïve Bayes neural classifier algorithm to classify the extracted texts and lexicon based polarity check will be perform thereafter. Naive Bayes Neural Classifier isused to classify sentence in more than one dimension. For example, the mobile is good and its music is superb. It covers two dimensions. The first one is mobile and the second one is music. It provides a strong impact on the sentiment words which are required for classifying the reviews.

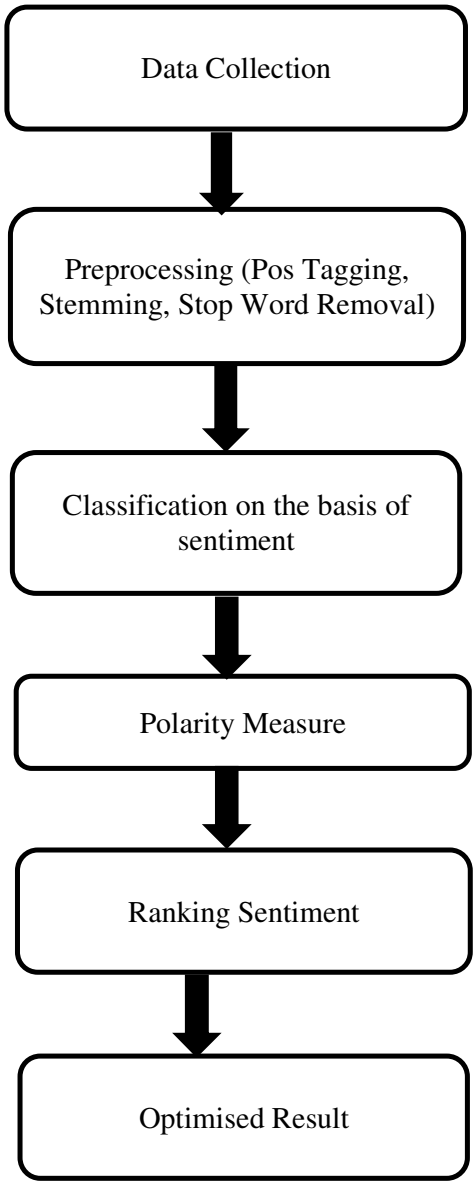

fig-1

After classifying the data in well manner, lexicon based polarity check will be perform. Typically, polarity check means evoking something positive or something negative when it is taken out of context. For example, the word beautifulhas a positive polarity, and the word ugly has a negative polarity. For polarity checking purpose we are going to use "sentiword.net" repository. After measuring the polarity of all the words, they will be rank accordingly, corresponding to their originateddocument. After ranking them all the optimized result will be obtained.

IV. CONCLUSIONS

It is seen that sentiment analysis/opinion mining play vital role to make decision about product /services.

Opinion Mining combines information retrieval and computational linguistic techniques handling an opinion on products, and review sites. Sentiment analysis deals with evaluation opinions or opinions type implying positive or negative sentiments. In this paper, we study some aspect and technique of sentiment analysis on products reviews and we approach some steps to get optimized result from those reviews.

REFERENCES

[1] Zheng-Jun Zha, Member, IEEE, Jianxing Yu, Jinhui Tang, Member, IEEE, Meng Wang, Member, IEEE, and Tat-SengChua,*Product Aspect Ranking and Its Applications.*

[2] GeorgLackermair, Daniel Kailer, KenanKanmaz, *Importance of Online Product Reviews from a Consumer's Perspective.*

[3] Minqing Hu and Bing Liu Department of Computer Science University of Illinois at Chicago, *Mining and Summarizing Customer Reviews.*

[4] DrS Vijayarani, Ms J Ilamathi,Ms. Nithya,*Preprocessing technique for text mining-An overview.*

[5] Theresa Wilson, JanyceWiebe, Paul Hoffmann, *Recognizing Contextual Polarity in Phrase-level Sentiment analysis.*

[6] WalaaMedhat, Ahmed Hassan, HodaKorashy, *Sentiment analysis algorithms and applications: A survey.*

[7] Lina L. Dhande, Dr. Prof. Girish K Patnaik,*Analysing sentiments of movie review data using naïve bayes neural classifier.*

[8] ChetanKaushik and Atul Mishra Computer Engg. Department, YMCA University of Science & Technology, Faridabad, *A Scalable, Lexicon Based Technique For Sentiment Analysis.*

[9] Ms. Gaurangi Patil1, Ms. Varsha Galande2, Mr. VedantKekan3, Ms. Kalpana Dange4, *Sentiment Analysis Using Support Vector Machine.*

[10] Haiyang Sui,Christopher Khoo,Syin Chan, *Sentiment Classification of Product Review using SVM and Decision Tree Induction.*

[11] Vijay Murari T 1, Sunil 2, Suchitra V3, Asst. Professor, Dept. of Computer Science, NMAMIT, Nitte, India, *Survey on Product Recommendation Using Customer Reviews Based on Opinion Mining.*

[12] S. Chandra Kala and C. Sindhu*Department of Computer Science and Engineering, Velammal Engineering College, India, Opinion Mining and Sentiment Classification: A Survey.*

[13] Anascollomb, CrinaCostea, DamienJoyeux, OmarHasan, LionelBrunie,*Study and Comparision of Sentiment Analysis methods for Reputation Evaluation.*

[14] MaiteTaboada,Julian Brooke, Milan Tofiloski, Kimberly Voll, Manfred Stede*, Lexicon-Based Methods for Sentiment Analysis*.

AUTHORS PROFILE:

Mrs.PinkySaikiaDutta, BE from Jorhat Engineering College, Assam and ME from Gauhati University in 2002 and 2014 respectively. She is currently working in Girijananda Chowdhury Institute of Management and Technology, Guwahati, Assam, India. She is a member of IEEE since from 2015. Her main research work focuses on Data mining, its algorithm. She has 13 years of teaching Experience.

Avishek Saha, pursuing B.Tech in Computer Science and Engineering deparment from in Girijananda Chowdhury Institute of Management and Technology, Guwahati, Assam,India. His area of interest is in Data Mining.

Jugal Kishor Talukdar, pursuing B.Tech in Computer Science and Engineering deparment from in Girijananda Chowdhury Institute of Management and Technology,Guwahati,Assam,India. His area of interest is in Data Mining

Amarendra Sharma, pursuing B.Tech in Computer Science and Engineering deparment from in Girijananda Chowdhury Institute of Management and Technology, Guwahati, Assam,India. His area of interest is in Data Mining.

Abhishek Sarmah, pursuing B.Tech in Computer /science and Engineering deparment from in Girijananda Chowdhury Institute of Management an Technology,Guwahati,Assam,India. His area of interest is in Data Mining.

.

