A Survey on Aspect based Opinion Mining

Pooja A. Rangari , Prof. Kalyani C. Waghmare

Abstract— **With the growth of internet social networking sites, blogs, forms have gained a tremendous importance. People comment on these sites to express their opinion. This information is of great importance for mining useful information from the text which can be done through opinion mining. Opinion mining or sentiment analysis is the computational field of study of people's opinions, emotions, and attitude towards particular aspect. In this paper, we are going to study in depth opinion mining and survey the existing methods that are being used for opinion mining.**

Index Terms— **Aspect, Opinion Mining, Sentiment Analysis, Topic Model.**

I. INTRODUCTION

With the advent of web 2.0, several types of social media sites such as blogs, discussion forums, review websites community websites and online shopping sites have emerged that have proved to be useful in determining the public; sentiment and opinion, towards the particular aspects of the products.

Thus with the growing use of Internet, use of social media sites have been increased to a much much larger extent. Even these days' online shopping websites have a much larger gained speed as people mostly prefer these sites for shopping. There are various merchants offering millions of products online. For example, 5 millions of products have been indexed by Bing Shopping. Amazon.com archives a total of more than 36 million products. Shopper.com records more than five million products from over 3,000 merchants [3].

Besides the retail Websites, platform for consumers to post reviews on millions of products is provided by many forum websites. For example, CNet.com involves more than seven million product reviews; whereas Pricegrabber.com contains millions of reviews on more than 32 million products in 20 distinct categories over 11,000 merchants. Such numerous reviews are very useful for mining useful information from the subjective for the benefit of Sellers and firms [4].

II. OPINION MINING

Opinion mining often referred to as Sentiment Analysis is the computational field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards particular entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. The major rich source of review feedback comments are blogs, forums, social networking sites, online shopping website, etc. People mostly comment on these sites from which users and buyers

Machine learning approaches are both supervised as well as unsupervised. But, opinion mining is basically a Supervised Approach where we need to train a classifier on the training set before it is to be applied on a test set. It combines the techniques of natural language processing, information retrieval, text analytics and computational linguistics.

Information gathering is an essential phase before taking a decision. Online review form sites, and personal blogs, tweets from twitter facilitate gathering of sentiments of products or object using information technologies. One of the main objective of Opinion mining is determining the polarity of comments whether they express positive, negative or neutral opinions by extracting features and components of the object that have been commented in each document.

Opinion mining is carried out at document level, sentence level, phrase level and aspect level mining based on the kind of information that is to be extracted from the opinionated text.

Every review consists of some attributes that must be considered while studying any review:

Opinion targets: Entities and their features/aspects
Sentiments: positive and negative
Opinion holders: persons who hold the opinions
Time: when opinions are expressed

III. MOTIVATIONAL SURVEY

The motivation behind this survey is that lots of information is been generated on the web these days. While buying a new product people mostly prefer to refer the opinions of other people who have bought the product and used it. Thus, for mining these essential facts from the reviews and to provide a short and fruitful output from a complete review opinion mining is useful [1].

Generally while commenting on a product people comments mostly about the aspects of the product. Identifying important product aspects will improve usability of numerous reviews and is beneficial to both consumers and firms [3]. Being a consumer, people mostly refer to the ratings of the product while buying a new product. Consumers can conveniently make wise purchasing decision by paying more attentions to the important aspects, while firms can focus on improving the quality of these aspects and thus enhance product reputation effectively.

IV. LITERATURE SURVEY

A. Sentiment Classification

Sentiment classification aims to classify the text in the document based on its polarity whether the text is having positive, negative or neutral polarity towards a particular

Manuscript received Feb 28, 2014.
Pooja A. Rangari*, Computer Engineering Department, Pune Institute of Computer Technology, Savitribai Phule Pune University., Pune, India.*
Prof. K. C. Waghmare*, Computer Engineering Department, Pune Institute of Computer Technology, Savitribai Phule Pune University., Pune, India.*

ISSN: 2278 – 7798 **All Rights Reserved © 2015 IJSETR**

aspect. According to (Go et al., 2009), it is not clear which of these classification strategies is the more appropriate to perform sentiment analysis. Hence, we study all the approaches.

The sentiment classification could be basically done via two approaches:

1. Machine Learning Approach
2. Lexicon based approach

1) Machine Learning Approach

In 2002 Bo Pang , Lee and Vaidyanathan [8] have suggested some classification techniques after doing implementation on movie review dataset. Machine Learning Approaches are best text categorization techniques [8] [9].

Lee et al. proposed that Naïve Bayes classifier is a probabilistic classifier which uses Bayes theorem [8]. Based upon the priori probability which is being obtained during training is used to build model which provides the formula for posterior probability, to calculate the object belongs to the result classes, and then make decision that posterior probability with maximum probability will give result class [10]. Naive Bayes-based text categorization still tends to perform surprisingly well (Lewis, 1998); indeed, Domingos and Pazzani (1997) show that Naive Bayes is optimal for certain problem classes with highly dependent features[8].

Lee et al. also proposed that Support Vector Machine is a supervised learning method widely used and best suited for text categorization. SVM is a binary classifier. The text to be classified is basically converted into word vectors. Then the hyper-plane is being drawn using these word vectors to separate the data instances of one class from another. SVM finds this hyper-plane using training instances also called support vectors [9].

Sheng Yuk et al. proposed that decision tree is a technique of data mining and machine learning. In this algorithm while travelling from root node to leaf, one entity will get the prediction result. Classification tree and regression tree are two basic and major types of decision trees. Classification tree analysis is applied when the prediction output is discrete classes. And regression tree is used when predicted outcome is continuous value [16].

Lim proposed a method which improves performance of kNN based text classification by using well estimated parameters [17]. Some variants of the kNN method with different decision functions, k values, and feature sets were proposed and evaluated to find out adequate parameters [18].

Anjarie et al. found that using SVM the efficiency of sentiment classifier have raised to 80% which was better than NaIve Bayes (NB) and decision tree [19]. Peng Zhang et al. applied RLSC to various dataset and commented on its efficiency comparable with SVM [6].

Weka 3.7 is a open source java tool .It is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java. Weka has inbuilt API's for naive bayes , SVM, J48 decision tree classifiers which could be used for text classification [21].

2) Lexicon based approach

The lexicon Approach predicts sentiment of review text using databases which contain word polarity values e.g. SentiWordNet. SENTIWORDNET 3.0, an enhanced lexical resource explicitly devised for supporting sentiment classification and opinion mining applications (Pang and Lee, 2008).Automatic annotation process according to which SENTIWORDNET 3.0 is generated. This process consists of two steps, (1) a weak-supervision, semi-supervised learning step, and (2) a random-walk step [11],[12].The SentiWordNet 3.0 dictionary consists of numerous number of synset terms followed by its glossary and the positivity, negativity and objectivity score for each term where, ObjScore = 1 - (PosScore + NegScore).

An equation that calculate sentiment polarity of review text using a resource [13] is

$$S(D) = \frac{\sum w \in D \; Sw.weight(w).modified(w)}{\sum weight(w)}$$

Where, S_w is a polarity score value of word *w*. This value is generated for dictionary using function weight (). Issues such as negation handling, word position in sentence etc are handled by operator modifier ( ).

B. Topic Modeling

Topic modeling provides a simple way to analyze large volumes of unlabeled text. Topic modeling captures word frequencies and co-occurrences effectively. A "topic" consists of a cluster of words that frequently occur together. Topic models can connect words with similar meanings and distinguish between uses of words with multiple meanings.

Firstly, the identification of salient topics needs to be done which are the aspects of the subject of the review. In this section, we divide all the topics based on its aspects. For example, we have mobile reviews and it consists of several aspects like battery, screen resolution, price, camera quality, RAM and memory which are the topics of the review. So, here in this step we apply topic modeling and separate all the reviews those are related to this topic in order to get the products reviews. Thus, we can use topic modeling approaches here for the automatic identification of features and depending on the prior knowledge of the topics, each topic works as a feature or one of semantic orientations [22].

T Landauer proposed the Latent Semantic Analysis a method for extracting and representing the contextual meaning of words by statistical computations applied to logare corpus text (Landauer and Dumais, 1997) [20]. LSA is a term document co-occurrence matrix where terms represent rows and each column stands for documents. LSA applies singular matrix decomposition (SVD) to matrix.

David Blei et al. proposed latent dirichlet allocation algorithm, a generative probabilistic model for collections of discrete data such as text corpora which was the first of its kind [14]. LDA is a three-level hierarchical Bayesian model, in which every item of the collection is modeled as a finite mixture over an underlying set of topics. LDA is thus used for topic modeling of text corpus.

Shenghua Bao et al. proposed Latent Dirichlet Allocation for Emotion Topic Modeling for selecting related documents based on their emotional preferences [2]. Hence, each document was classified among 16 emotions. The dataset used was news review dataset sina.com one of the largest news portal in China. Also, S. Sujitha et al. proposed LDA

ISSN: 2278 – 7798 All Rights Reserved © 2015 IJSETR

with an extra layer of emotion topic modeling [15]. Here, an Emotion Term model was used which was fused with Topic Model. During the research, it was believed that the human summarize their emotion of the overall document by the last sentence of the document.

Xiuzhen Zhang et al. proposed a multi dimensional trust model for computing reputation trust scores for sellers from user feedback comments and using Lexical LDA for dimension ratings and weight [1]. The dataset used was eBay free text feedback comments on sellers. Detailed seller ratings for sellers were mainly based on four aspects namely item, communication, postage time, handling charges. As for now, the score for each seller is the same for each of his product. The system consisted that the all good reputation trust score for each seller was aggregated to be different for each item that the seller sells which is the first piece of work on trust evaluation by mining feedback comments [14].

C. *Aspect Identification*

Aspect is defined as the subject of review on which user comments such as for a product like "mobile" its aspects would be "battery", "display", "camera", etc. Aspects are divided into two type's implicit as well as explicit aspects [23]. By explicit it means that aspect itself occurs in the sentence for example "The camera of this phone is good quality" here the sentence clearly defines about the good camera quality of phone. But, the sentence "It's too large to handle" does not give a clear view that it is describing about the large size of the camera.

Hu and Liu firstly proposed the method to extract product aspects based on association rule mining. The idea behind the research was that the consumer tends to use the same words when they comments on the same products aspects, and then frequent item sets of nouns in reviews are likely to be the products aspects . In this system, the occurrence frequencies of noun and noun phrases where counted and only the frequent nouns are kept as aspects [24].

Qiu et al. proposed the idea that opinion words can be used to detect product aspects and vice versa, focusing on single reviews. In this approach, a seed set of opinion words is combined with syntactic dependencies to identify product aspects and new opinion words [25]. To detect the polarity of the newly identified opinion words, they consider the given polarities of the seed words and make the assumption that opinion words expressing a sentiment towards the same aspect in the same review share the same polarity.

Zhu et al. used product aspects and aspect-related terms as input for their algorithm, and aimed to discover new aspect-related terms by applying a bootstrapping algorithm based on co-occurrence between seed terms and new candidate terms. A sentiment score is again obtained by accessing an external sentiment lexicon [7].

Zheng Zha et al. identified the aspects extracting the frequent noun terms in the reviews. For identifying aspects in the free text reviews, a straightforward solution is to employ an existing aspect identification approach. They followed the approach proposed by hu and Liu that the occurrence frequencies of the nouns and noun phrases are counted, and only the frequent ones are kept as aspects [3].

V. CONCLUSION

Due to social networking sites, large amount of data is being generated on internet every second. This data is of rich source for mining useful information from the text. Opinion mining Analysis is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards particular entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. In this paper, we have seen in detailed about aspect based opinion mining and techniques for sentiment classification, topic modeling and aspect identification.

REFERENCES

[1] Xiuzhen Zhang, Lishan Cui, Yan Wang, "CommTrust: Computing Multi Dimensional Trust by Mining E-Commerce Feedback Comments," IEEE Transactions On Knowledge and Data Engineering., vol. 26, No. 7, July 2014.

[2] Shenghua Bao , Shengliang Xu , Li Zhang, Rong Yan, Dingyi Han, and Yong Yu, "Mining Social Emotions from Affective Text," IEEE Transactions On Knowledge and Data Engineering, Vol. 24, No . 9, September 2012.

[3] Zheng-Jun Zha, *Member, IEEE*, Jianxing Yu, Jinhui Tang, *Member, IEEE,* Meng Wang, *Member, IEEE*, and Tat-Seng Chua "Product Aspect Ranking and Its Applications" Ieee Transactions On Knowledge And Data Engineering, Vol. 26, No. 5, May 2014

[4] A. Ghose and P. G. Ipeirotis, "Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics", IEEE Trans. Knowl. Data Eng., vol. 23, no. 10, pp. 1498–1512. Sept. 2010.

[5] " A survey on opinion mining and sentiment analysis" Bing Liu and Lei Zhang Nov 2010.

[6] Malhar Anjaria, Ram Mahana Reddy Guddeti ," Influence Factor Based Opinion Mining of Twitter Data Using Supervised Learning" , IEEE 2014.

[7] Zhu J., Wang H., Zhu M., Tsou B. K., Ma M., "Aspect-based opinion polling from customer reviews", IEEE Transactions on Affective Computing, 2(1):37–49, 2011.

[8] Bo Pang and Lillian Lee, Shivkumar Vaidyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques", Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002).

[9] N. Anitha, B. Anitha, S. Pradeepa, "Sentiment Classification Approaches – A Review", International Journal of Innovations in Engineering and Technology (IJIET) Vol. 3 Issue 1 October 2013.

[10] Vivek Narayanan, Ishan Arora, Arjun Bhatia, "Fast And Accurate Sentiment Classification Using An Enhanced Naive Bayes Model", 14th International Conference, IDEAL 2013, Hefei, China, October 20-23, 2013. Proceedings , pp 194-201.

[11] Esuli, A., & Sebastiani, F. (2006). "SentiWordNet: A publicly available resource for opinion mining". In Proceedings of the 6th international conference on Language Resources and Evaluation (LREC'06), pp.417–422.

[12] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani, "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining", 2010.

[13] Mikalai Tsytsarau, Themis Palpanas, "Survey on mining subjective data on the web", Data Mining Knowledge Discovery, Springer 2012, pp.478-514.

[14] DM Blei, AY Ng, MI Jordan, "Latent Dirichlet Allocation," The Journal of machine Learning research 3, pp. 993-1022 , 2003.

[15] S.Sujitha, S.Selvi, J.Martina Jasmine, "Emotion Classification of Textual Document Using Emotion Topic Model," International Journal of Advanced Research in Computer Engineering and Technology (IJARCET), Volume 3, Issue 2, February 2014.

[16] Sheng Yu, Subhash Kak "A Survey of Prediction Using Social Media" Social and Information Networks *7 Mar 2012.*

[17] Heui Lim, Improving kNN Based Text Classification with Well Estimated Parameters, LNCS, Vol. 3316, Oct 2004, Pages 516 - 523.

[18] M. Ikonomakis, S. Kotsiantis, V. Tampakas " Text Classification Using Machine Learning Techniques" WSEAS TRANSACTIONS on COMPUTERS, Issue 8, Volume 4, August 2005, pp. 966-974

[19] Thomas K Landauer , Peter W. Foltz , Darrell Laham, "An Introduction to Latent Semantic Analysis," Discourse Processes, 25, 259-284, 1998.

ISSN: 2278 – 7798 All Rights Reserved © 2015 IJSETR

[20] Qiu G., Liu B., Bu J., Chen C., Expanding domain sentiment lexicon through double propagation. In Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence, volume 9, pages 1199–1204, 2009.
[21] Peng Zhang and Jing Peng , "SVM vs Regularized Least Squares Classification", Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04).
[22] http://www.cs.waikato.ac.nz/ml/weka/
[23] Hyun Duk Kim, Kavita Ganesan, Parikshit Sondhi, Chengxiang Zhai, "Comprehensive Review of Opinion Summarization"
[24] Mily Lal. Kavita Asnani"Aspect Extraction & Segmentation In Opinion Mining", International Journel of Engineering and Computer Science" ISSN:2319-7242 Volume 3 Issue 5 May, 2014 Page No. 5873-5878.
[25] Hu M and Liu, "Mining and Summarizing customer reviews" International conference on knowledge and data mining 2004.

Miss. Pooja A. Rangari received B.E degree in Computer Science and Engineering from Sipna College of Engineering and Technology, Amravati and currently pursuing her M.E degree in Computer Engineering from Pune Institute of Computer Technology, Pune.

Prof. Kalyani C Waghmare, is Assistant Professor in Computer Engineering department at Pune Institute Of Computer Engineering, Pune. She has major areas of interest in data mining, data structures, Design & analysis of Algorithm.

ISSN: 2278 – 7798 All Rights Reserved © 2015 IJSETR