ISSN No. 0976-5697

Volume 4, No. 8, May-June 2013

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Opinion Mining Task and Techniques: A Survey

Harpreet Kaur
Assistant Professor
Department of Information Technology
DAV Institute of Engineering and Technology
Jalandhar, Punjab
sonia29590@gmail.com

Priya
(M.Tech Student)
Department of Computer Science and Engineering
DAV Institute of Engineering and Technology
Jalandhar, Punjab
priya.21690@gmail.com

Abstract: Opinion mining is a process of Tracking and extracting the knowledge of public about some products or problem. Opinion mining is an application of natural language processing to identify objective and subjective information from others. Earlier Questionnaire based surveys were conducted using traditional tools designed to collect information about user experience. Today online review sites and personal blogs are sources for opinion sharing. Opinion mining is also known as sentimental analysis as it represents user's sentiments, feelings or appraisal related to the same. In this paper we have presented techniques and methods to get opinion oriented information. It includes both subjective and objective phase of opinion mining. We have tried to recognize most casually used techniques and methods implemented for opinionated documents to assist our future research in this area.

Keywords: Opinion mining, Sentiment analysis, Opinion analysis, Machine learning.

I. INTRODUCTION

Opinion mining deals with the analysis of the sentiments of the people and accessing their opinions, emotions and attitudes. Opinion mining is an emerging area of research which extensively uses information retrieval and computational linguistics. Opinions of people always play a vital role in decision making process and are an essential part of effective learning. Opinions are widely shared on many web portals like review sites or pages, corporate websites, discussion groups, blogs, voting portals etc. apart from helping people to express their ideas and thoughts it also allow business organizations to endeavor and enhance their capabilities and products by providing best of their services according to the reasonable demands of the people. It deals with the knowledge discovery from text using Data mining and Natural Language Processing Techniques [1]. Earlier Questionnaire based surveys were conducted which were the Traditional tools and were used to extract opinions of people about various products and services. Business organizations conduct surveys to know consumers sentiments and opinions about their products.

These Types of surveys are easy to conduct and easy for collecting information. Surveys are objective and can be subjective. In objective surveys questions are asked along with options to select. In subjective Type surveys user need to write comments regarding products. Some surveys can be conducted using both the types that mean user can be asked to select one of the options as well as to provide the reasons as well. But sometimes user's comments can elaborate only about negative aspects while providing high score which will pair a negative text with a positive rating. Such a pairing will result in Wrong evaluation of results and thus results in an inaccurate model for classification. Therefore when such a model will be used for conducting surveys then it might lead to some problems [3].The possible solution to this problem can be the integration of all the comments of the users into one segment as whole. This complements feature based opinion analysis. Thus it focuses on the extraction of the text documents as whole rather than focusing on particular attributes of the products. So the negative and positive opinions of the consumers will be integrated together so that the decision making will result accurate and efficient.

Data on WWW is dynamic in nature as it changes rapidly due to continuous updating and addition of latest information .Opinion mining can be easily illustrated by taking daily routine examples. When a person wants to buy some product, the person consults his friends, relatives and neighbors about the brand, quality, and variety and then buys a product. Most of us get help from others. It was done verbally or through letters. We ask for the suggestions regarding products that whether the product is good or bad, about its quality etc. But now it is easy to share knowledge as the number of blogs and survey sites are available that collect opinions about the products. Spam filtering is a process of detection and removal of fake opinions. This misleads the users by giving unworthy positive or negative opinions to some objects. It is also a research issue in healthy opinion mining [5].

Questions in the surveys are regarding the attributes, services, quality and quantity of products. Opinion classification has been studied by the natural language processing community [5,6] and defines the following: Suppose a set of text data D is given, the classification analyzes whether each document $d \in D$ expresses a positive or negative opinion about certain object. For example, given a set of blogs on cosmetics reviews, positive reviews and negative reviews will be classified by the system. In opinion classification opinion words that indicate positive or negative opinions are important, e.g., great, excellent, amazing, horrible, bad, worst, etc. Opinions of the consumers can be positive or negative. Opinions about the products are stored or collected and

according to that decision making process is carried out equally by manufacturer, customer and merchant. Most of the methodologies for opinion mining apply some forms of machine learning techniques for classification.Fig.1 represents an example of user rating of a restaurant that describes whether the services, cleanliness, parking, quality of food and choice of food of the restaurant are very poor, poor, OK, good or very good.

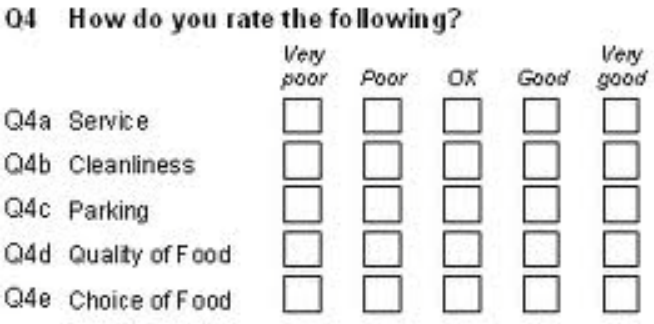

Figure 1

II. TASKS WITHIN OPINION MINING AND SENTIMENT ANALYSIS

Usually people think that opinion mining is all about to predict the polarity of the text as positive or negative. But it's not true. It is much broader task. The tasks are not popular that's why people are not aware about them. The Tasks included are:

A. Subjectivity Detection:

This task includes whether the text contains opinions or not (subjective or objective).

B. Sentiment Prediction:

This task includes the prediction of polarity of the text i.e. whether the text is positive or negative. It can be at document level, sentence level and at phrase level.

C. Aspect Based Sentiment Summarization:

This task includes the discovery of features of products and then discovering the sentiments for each feature. Example-for HTC mobile you may have features like design, sound, screen etc.

D. Contrastive Viewpoint Summarization:

This task includes the contradicting opinions .For Example-some may like the policies of some particular company and some may not. So this creates contrastive viewpoints of people.

E. Text Summarization for Opinions

This task includes the useful summary format to generate textual summaries .Example: a few sentences summarizing the reviews of a product.

F. Product Feature Extraction:

This task includes the extraction of features of products. It includes the following steps:

a. Identify product features.
b. Identify opinions regarding product features.
c. Determine the polarity of opinions.
d. Rank opinions based on their strength.

G. Opinion Retrieval:

This task includes the retrieval of documents and ranking according to their opinions about a query topic. The document must satisfy two criteria's: It should be relevant to the query topic and should contain opinions about the query.

H. Opinion-Based Entity Ranking:

All the candidate entities are ranked that are based on how well opinions on these entities match the user's preferences.

III. ASSOCIATED RESOURCES

This section depicts some necessary resources taken under consideration which are used in summarizing the opinions in order to aid the extraction of opinions from text [7]. This process mainly involves two steps: firstly extracting the feature words from the given review and then producing a summary by assigning scores to the feature extracted words from the first step. Some available lexical resources for opinion mining are:

a. ***Senti Word Net:*** It is a lexical resource in which each WordNet is associated to three numerical scores Obj, Pos, and Neg, describing how objective, positive and negative the terms are present in the review. Each term in SentiWordwNet is associated with numerical scores for positive and negative sentiment information. In this resource, the analysis depends upon two polarities. SO-Polarity which helps in deciding whether a given text has a factual nature (i.e. describes a given situation or event, without expressing a positive or a negative opinion on it) or expresses an opinion on its subject matter. PN-Polarity helps in deciding if a given subjective text expresses a positive or a negative opinion on its subject matter. Figure 2 shows the graphical model designed to display the scores in this resource.

b. ***Stanford POS Tagger:*** Parts-of-Speech tagging is a process of assigning tags such as noun, adjective and adverbs to different parts of speech whereas a Part-of-Speech Tagger is a piece of software that reads text in some language, processes

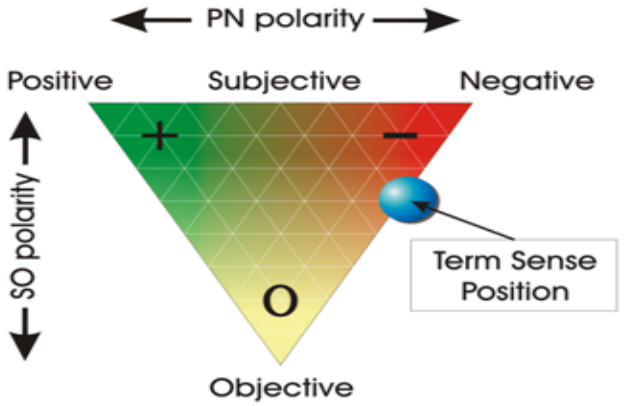

Figure.2: The graphical representation adopted by SENTIWORDNET for representing the opinion-related properties of a term sense.

A sequence of words and label it automatically as part of speech to each word such as noun, verb, adjective etc. The English taggers use the Penn Tree bank tag set. Some key examples are:

a) NN: Singular noun
b) NNP: Proper singular noun
c) NNS: Plural noun
d) RB: Adverb
e) PRP: Personal pronoun
f) VB: Verb, base form
g) JJ: Adjective
h) NNPS: Proper plural noun
i) RRB: Right bracket
j) LRB: Left bracket

These mark up a word in a text corresponding to a particular part of speech, based on both its definition, as well as its context.

*c. **Word Sense Disambiguation:*** The methods or resources described above are based on lexicon resources with the help of which it is unable to identify the true sense of the word due to which WSD has overshadowed these methods. WSD i.e. Word Sense Disambiguation is defined as the process of finding the sense of a word in a sentence especially when the same word is representing different meanings [9]. Usually this process firstly encompass WordNet glossaries which extract all possible patterns for all different opinion expressions and then the senses are ranked according to their relevance to the text.

Prior to WSD pre-processing of text is required [10] which removes the unusable and incoreect data from the opinionated text such as non-textual context, date of comment, name of the reviewer etc. Sentiment analysis with the help of WSD involves number of steps which include Parts of speech tagging and classifying the polarity of the words using SentiWordNet technique. After evaluating these words if there exists mutiple senses of the same word then WSD classify the occurrence of the word in context into one or more of its sense classes and then polarities like positive, negative and neutral are assigned to them.

Major areas involving the applications of WSD are Machine translation, information retrieval, lexicography, knowledge mining and is becoming increasingly important in new research area such as bio-informatics and the Semantic web.

IV. MACHINE LEARNING METHODS

Machine learning focuses on prediction, based on known properties learned from the training data. In order to train a classifier for sentiment recognition in text, classic supervised learning techniques (e.g. Support Vector Machines, naive Bayes, Maximum Entropy) can be used [6]. A supervised approach entails the use of a labeled training corpus to learn a certain classification function. The method that in the literature often yields the highest accuracy regards a Support Vector Machine classifier.

A. The Naive Bayes Classifier:

The Naive Bayes classifier is based on Bayesian probability and assumed that feature probabilities are independent of one another. To assign to a given document d the class C^* =arg max_c $P(c \mid d)$.We derive the Naive Bayes (NB) classifier by first observing that by Bayes rule,

$$P(c|d)=\frac{P(c)P(d|c)}{P(d)}$$

Where P (d) plays no role in selecting c^* .To estimate the term P (d | c), Naive Bayes decomposes it by assuming the f_i's are conditionally independent given

$$P(c|d) := \frac{P(c)(\pi_{i=1}^{m} P(f_i|c)^{n_i(d)})}{P(d)}$$

Our training method consists of relative-frequency estimation of $P(c)$ and $P(fi \mid c)$ using add-one Smoothing .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. On the other hand, more sophisticated algorithms might (and often do) yield better results. The Naive Bayes classifier seems to be simple, it is observed to have high predictive power; in our tests, it performed competitively with the more sophisticated classifiers we used. The Bayes classifier can also be implemented very efficiently. Its independence assumption means that it does not fall prey to the curse of dimensionality, and its running time is linear in the size of the input.

B. Maximum Entropy Classifier:

This machine learning technique provides the least biased estimate possible based on the given information. It makes no conditional independence assumption between features, as the Naive Bayes classifier does. Maximum entropy's estimate of takes the following exponential form:

$$P(c|d)=\frac{1}{Z(d)} \exp\left(\sum_i(\lambda_{i,c}\ F_{i,c}\ (d,c))\right)$$

The $\lambda_{i,c}$'s are feature-weigh parameters, where a large $\lambda_{i,c}$ means that F_i is considered a strong indicator for class c. Pang used the Improved Iterative Scaling (IIS) method, but L-BFGS, a method that was invented after their paper was published, was found to out-perform IIS and generalized iterative scaling (GIS), yet another parameter estimation method.

C. The Support Vector Machine Classifier:

Support Vector Machines (SVMs) operate by separating points in a d-dimensional space using a (d-1)-dimensional hyperplane. Given a set of training data, the SVM classifier finds a hyperplane with the largest possible margin; that is, it tries finds the hyperplane such that each training point is correctly classified and the hyperplane is as far as possible from the points closest to it. In practice, it is usually not possible to find a hyperplane that separates the classes perfectly, so points are permitted to be inside the margin or on the wrong side of the hyperplane. Any point on or inside the

margin is referred to as a support vector, is selected through a constrained quadratic optimization to minimize

$$\frac{1}{2} \left|\vec{B}\right|^2 + C\sum_i \delta_i$$

$$\forall i, \delta_i \geq 0$$

$$\forall i, y_i(\vec{x_i^T}.\vec{B} + B_0) \geq 1 - \delta_i$$

V. APPLICATIONS

The applications of opinion mining are in

a. Argument mapping software helps organizing in a logical way these policy statements, by explicating the logical links between them.
b. Politics: As is well known, opinions matter a great deal in politics. Some work has focused on understanding what voters are thinking.
c. Automated content analysis helps processing large amount of qualitative data [5].
d. Applications in Business marketing intelligence, product and service benchmarking and improvement. To understand the voice of the customer as expressed in everyday communications.

VI. OPINION MINING AT DIFFERENT LEVELS

There are three types of opinion mining [6]. First one is Document Level opinion mining in which [6] the whole document is written about only one product and contains opinion posted by a single opinion holder. Next is Feature Level opinion mining [6], in the data source focuses on features of a single object posted by single opinion holder. All the features or attributes are separated and for particular feature the opinions are extracted. Both of the techniques are complicated. And the last is Sentence Level opinion mining [6], in which Different people who have already used product, have written their opinions for product. A sentence contains only one opinion posted by single opinion holder; this could not be true in many cases e.g. there could be multiple opinions in compound.

There are three techniques to used Naïve Bayesian algorithm [8]. The first one is Machine Learning[6] , in this Natural Language Processing algorithm are used Next is Semantic Analysis Pattern based, in which co relations between the words of the sentence are found.. Last one is Term Counting based[8], in which the number of negative and positive words are count from the sentence and if more number of negative words, then the opinion is negative and if more number of positive words, then opinion is positive. If the dictionary or database is good then it really gives good results.

According to Supervised Term Counting based approach of Naïve Bayesian algorithm [8] the probabilities of the labels, according to the words are found. Two algorithms original naïve Bayesian algorithm and modified Naïve Bayesian algorithm are used. The accuracy of modified naïve Bayesian algorithm is more than original Naïve Bayesian algorithm.

Table 1. Opinion mining at different levels[5]

Classification of Opinion mining at different levels	Assumptions made at different levels
1. Opinion Mining at Document level	The whole document is written about only one product and contains opinion posted by a single opinion holder. The results are presented using naive bayes, maximum entropy and support vector machine algorithms and good results are shown comparable to other ranging from 71 to 85% depending on the method and test data sets.
2. Opinion Mining at Sentence level	Different people, who have already used product, have written their opinions for product. A sentence contains only one opinion posted by single opinion holder; this could not be true in many cases e.g. there could be multiple opinions in compound. Secondly the sentence boundary is defined in the given document
3. Opinion Mining at Feature level	The data source focuses on features of a single object posted by single opinion holder. All the features or attributes are separated and for particular feature the opinions are extracted.

VII. CONCLUSION

As Opinion mining is the emerging and rapidly growing research area, our goal in this paper has been to cover mainly on the crucial methods and techniques to explore the domain in such topics. This paper generally focuses on the tasks, resources and techniques involved in opinion mining and sentiment analysis along with some machine models used to implement it. Further work on this area will result in more enhancements in the short term and long term aspects of our future research. Although a lot of research has been performed in past on its wide area but due to lack of availability of corpus we mainly concentrated on its lexical resource and machine learning techniques. We also tried to explore various tasks with the help of which opinions can be extracted from the sources over supervised learning. This survey is based upon the work supported by various authors but we tried to integrate the relevant issues related to opinion extraction of the people from various resources which help in the aid of extracting the intensities and polarities of likes and dislikes of people. There are many improvements which can be made to the opinion mining application in terms of making use of further linguistic and contextual clues. So, the survey conducted by us is hopefully of significant use to help researchers in knowing about the past and recent trends in same area.

VIII.REFERENCES

[1] Khairullah Khan, Baharum B.Baharudin, Aurangzeb Khan, Fazal-e-Malik, "Mining Opinion from Text Documents", Proceedings of 2009 3rd IEEE International Conference on Digital Ecosystems and Technologies.

[2] Anshu Jain,Pragya Shukla,Hitesh Bandiya, ANALYZING SENTIMENTS IN PRODUCT REVIEWS, IJRIM Volume 2, Issue 2 (February 2012) (ISSN 2231-4334).

[3] Evgeny A. Stepanov, Giuseppe Riccardi, "Detecting General Opinions from Customer Surveys", Proceedings of 11th IEEE International Conference on Data Mining Workshops, 2011.

[4] Gokul Patil1, Amit Patil, "Web Information Extraction and classification using Vector Space Model Algorithm", International Journal of Emerging Technology and Advanced Engineering,ISSN 2250-2459, Volume 1, Issue 2, December 2011.

[5] Nidhi Mishra, C.K.Jha, PhD, Classification of Opinion Mining Techniques, International Journal of Computer Applications (0975 – 8887) Volume 56– No.13, October 2012.

[6] Bo Pang and Lillian Lee, Shivakumar Vaithyanathan, Thumbs up? Sentiment Classification using Machine Learning Techniques, Proceedings of EMNLP 2002, pp. 79–86.

[7] Andrea Esuli and Fabrizio Sebastiani." SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining", Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche via Giuseppe Moruzzi 1, 56124 Pisa, Italy.

[8] Trivedi Khushboo N, Swati K. Vekariya, Prof.Shailendra Mishra, Mining of Sentence Level Opinion Using Supervised Term Weighted Approach of Naïve Bayesian Algorithm, Vol 3 (3), 987-991, IJCTA | MAY-JUNE 2012.

[9] Mingqing Hu and Bing Liu,"Mining Opinion Features in Customer Reviews",American Association for Artificial Intelligence,2004.

[10] Aurangzeb khan, Baharum Baharudin, "Sentiment Classification using Sentence Level Semantic Orientation of Opinion Terms from Blogs",Proceedings of 2011 IEEE.