A Survey on Sentiment Analysis Algorithms for Opinion Mining

Vidisha M. Pradhan
M.E. Student
Department of Information Technology
GCET, V. V. Nagar, Affiliated to G.T.U.
Gujarat, India

Jay Vala
Assistant Professor
Department of Information Technology
GCET, V. V. Nagar, Affiliated to G.T.U.
Gujarat, India

Prem Balani
Assistant Professor
Department of Information Technology
GCET, V. V. Nagar, Affiliated to G.T.U.
Gujarat, India

ABSTRACT
Opinion mining and sentiment analysis is rapidly growing area. There are numerous e-commerce sites available on internet which provides options to users to give feedback about specific product. These feedbacks are very much helpful to both the individuals, who are willing to buy that product and the organizations. An accurate method for predicting sentiments could enable us, to extract opinions from the internet and predict customer's preferences. There are various algorithms available for opinion mining. Before applying any algorithm for polarity detection, pre-processing on feedback is carried out. From these pre-processed reviews opinion words and object on which opinion is generated are extracted and any opinion mining technique is applied to find the polarity of the review. Opinion mining has three levels of granularities: Document level, Sentence level and Aspect level. In this paper various algorithms for sentiment analysis are studied and challenges and applications appear in this field are discussed.

Keywords
Sentiment Analysis, Opinion Mining, Web Content, Machine Learning.

1. INTRODUCTION
Sentiment Analysis (SA) or Opinion Mining (OM) is the computational study of people's opinions, attitudes and emotions toward an entity [3]. In general, opinion mining helps to collect information about the positive and negative aspects of a particular topic. Finally, the positive and highly scored opinions obtained about a particular product are recommended to the user. In order to promote marketing, large companies and business people are making use of opinion mining [4] .

Much research exists on sentiment analysis of user opinion data, which mainly judges the polarities of user reviews. In these studies, sentiment analysis is often conducted at one of the three levels: the document level, sentence level, or attribute level. In relation to sentiment analysis, the literature survey done indicates two types of techniques including machine learning and semantic orientation are important [3].These techniques are shown in figure 1.

There are several challenges in Sentiment analysis. The first is a opinion word that is considered to be positive in one situation may be considered negative in another situation. A second challenge is that people don't always express opinions in a same way. Most traditional text processing relies on the fact that small differences between two pieces of text don't change the meaning very much. In Sentiment analysis, however, "the picture was great" is very different from "the picture was not great". People can be contradictory in their statements [4]. Most reviews will have both positive and negative comments, which is somewhat manageable by analysing sentences one at a time.

Users express their opinions about products or services they consume in blog posts, shopping sites, or review sites. It is useful for both the consumers as well as for the producers to know what general public think about a particular product or service[6] . In the informal medium like twitter or blogs, the more likely people are to combine different opinions in the same sentence which is easy for a human to understand, but more difficult for a computer to parse. Sometimes even other people have difficulty understanding what someone thought based on a short piece of text because it lacks context [5] .

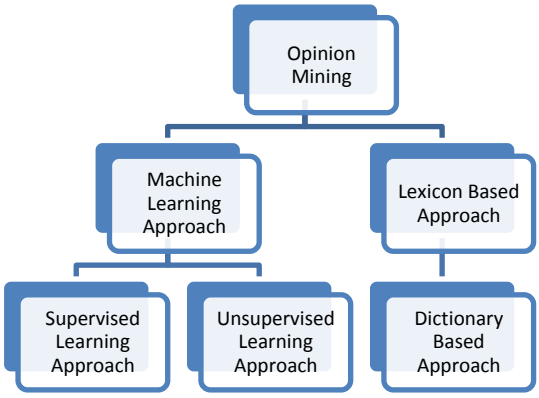

Figure 1. Opinion Mining Techniques

There are 3 levels of opinion mining:

1. Document Level
In this approaches whole document is considers as a single entity and the analysis approaches in applied on the whole document. The result generated in document level sometimes not appropriate [5,6].

2. Sentence Level
In the sentence level approaches every sentence is considered as an entity and analysis approaches is applied on individual sentence then their result is summarized to provide the overall result of the document [5,6].

3. Aspect Level
Phrase-level opinion mining is also known as aspect based opinion mining. It performs fine grained analysis and directly looks at the opinion. The goal of this level of analysis is to discover sentiments on aspects of items [5,6].

- Aspects that are explicitly mentioned as nouns or noun phrases in a sentence are called as **explicit aspects**.
- **Implicit aspects** are not explicitly mentioned in a sentence but are implied

There are different types of algorithms to analyze sentiments. In this paper all the techniques used for opinion mining is surveyed. Issues and applications of opinion mining are also discussed.

The rest of the paper is organized in the following way. Section 2 gives the overview of the techniques used for opinion mining with its related work. Section 3 contains issues in opinion mining. Section 4 contains applications of opinion mining . Section 4 contains conclusion.

2. RELATED WORK

2.1 Supervised Learning Approach

This method contains two sets of documents which are training and a test set. To learn about the document, training set is used by classifier. For validation purpose test set is used. For review classification many techniques can be used.

Types of supervised learning methods:

2.1.1 Decision tree classifier

Decision tree classifier provides a hierarchical decomposition of the training data space in which a condition on the attribute value is used to divide the data. The condition or predicate is the presence or absence of one or more words. The division of the data space is done recursively until the leaf nodes contain certain minimum numbers of records which are used for the purpose of classification.

In [7] Movie review features obtained from IMDb was extracted using inverse document frequency and the importance of the word found. Principal component analysis and CART were used for feature selection based on the importance of the work with respect to the entire document. The classification accuracy obtained by LVQ was 75%.

Exploring emotional variation in adolescent age and reasons behind these changes using data mining techniques is proposed in [11]. By classifying emotions and using decision tree different emotional variations are analyzed. If-then rules are also generated from decision tree. Outlier analysis is used to identify emotion variation in child having any kind of disability.

2.1.2 Linear classifier

a. Support vector machine:

Text data are ideally suited for SVM classification because of the sparse nature of text, in which few features are irrelevant, but they tend to be correlated with one another and generally organized into linearly separable categories.

In [10], machine learning (SVM) combined with domain specific lexicons is implemented for aspect classification and polarity identification of product review. SVM is trained to model aspect classification and this trained SVM is used for polarity classification per aspect. The experimental results indicate that the proposed techniques have achieved about 78% accuracy. Web based data are applied to emotion cause extraction sub system and complementary feature selection method, based on the output of these features are merged. In training process, web post with unknown emotions are given to SVM and SVR classification model and the output gives information about the type of emotion [13].

b. Neural network

Neural Network consists of many neurons where the neuron is its basic unit. The inputs to the neurons are denoted by the vector over line Xi which is the word frequencies in the ith document. There are a set of weights A which are associated with each neuron used in order to compute a function of its inputs. Based on inputs and weights output is generated.

2.1.3 Rule based classifier

In rule based classifiers, the data space is modeled with a set of rules. The left hand side represents a condition on the feature set expressed in disjunctive normal form while the right hand side is the class label. The conditions are on the term presence. Term absence is rarely used because it is not informative in sparse data.

[8] proposes a rule-based approach to emotion cause component detection for Chinese micro-blogs. It presents the emotion model and extracts the corresponding cause components in fine-grained emotions. The emotional lexicon can be constructed manually and automatically from the corpus. Meanwhile, the proportions of cause components can be calculated in the influence of the multi-language features based on Bayesian probability. The experiment results show the feasibility of the approach.

2.1.4 Probabilistic classifier

a. Naïve bayes

The Naive Bayes classifier is the simplest and most commonly used classifier. Naive Bayes classification model computes the posterior probability of a class, based on the distribution of the words in the document. The model works with the BOWs feature extraction which ignores the position of the word in the document. It uses Bayes Theorem to predict the probability that a given feature set belongs to a particular label.

The system which is proposed in [6] extracts aspects in product customer reviews. The nouns and noun phrases are extracted from each review sentence. Minimum support threshold is used to find all frequent aspects for the given review sentences. Naïve Bayesian algorithm using supervised term counting based approach is used to identify whether sentence is positive or negative opinion and also identifies the number of it.

The paper [12] presents a method of sentiment analysis, on the review made by users to movies. Classification of reviews in both positive and negative classes is done based on a naive Bayes algorithm. As training data we used a collection (pre-classified in positive and negative) of sentences taken from the movie reviews. To improve classification we removed insignificant words and introduced in classification groups of words (n-grams). For n = 2 groups we achieved a substantial improvement in classification.

b. Bayesian network

The main assumption of the NB classifier is the independence of the features. The other extreme assumption is to assume that all the features are fully dependent. This leads to the Bayesian Network model which is a directed acyclic graph whose nodes represent random variables, and edges represent conditional dependencies. BN is considered a complete model for the variables and their relationships. In Text mining, the computation complexity of BN is very expensive; that is why, it is not frequently used.

c. Maximum entropy

The Maximum entropy Classifier (known as a conditional exponential classifier) converts labeled feature sets to vectors using encoding. This encoded vector is then used to calculate weights for each feature that can then be combined to determine the most likely label for a feature set.

In [11], a novel method is used to collect various learners twitter messages On this dataset preprocessing for sentiment analysis is performed It involves various intermediate operations remove ambiguity. The pre-proccesed dataset is used to built user's emotional state classification and SVM, ME and naïve bayes classifiers are applied and the results are very efficient.

Table 1. Accuracy of Various Supervised Algorithm

Title, Author, Publication	Method	Accuracy
Title: Opinion Mining Using Decision Tree Based Feature Selection Through Manhattan Hierarchical Cluster Measure [7] **Author:** Jeevanandam Jotheeswaran, Dr. Y. S. Kumaraswamy **Publication:** Journal of Theoretical and Applied Information Technology, 2013	Naive bayes	Pricision: 0.50 Recall: o.79
Title: Sentiment Analysis: Measuring Opinions [10] **Author:** Chetashri Bhadane,Hardi Dalal, Heenal Doshi **Publication:** Science Direct, 2015	SVM	78%
Title: Analysis And Indentifying Variation In Human Emotion Through Data Mining [11] **Author:** Jasakaran Kaur, Sheveta Vashisht **Publication**: Int.J Computer Technology & Applications, 2012	Decision Tree	NA
Title: Twitter Sentiment Mining (Tsm) Framework Based Learners Emotional State Classification And Visualization For E-Learning System [9] Aut**hor:** M.Ravichandran, G.Kulanthaivel **Publication:** Journal of Theoretical and Applied Information Technology, 2014	SVM, Maximum Entropy	95%, 95%
Title: A Rule-Based Approach To Emotion Cause Detection For Chinese Micro-Blogs [8] **Author:** Kai Gao, Hua Xu, Jiushuo Wanga **Publication:** ELSEVIER, 2015	Assiciation rule	75%
Title: Extracting Aspects And Mining Opinions In Product Reviews Using Supervised Learning Algorithm [6] **Author**: A.Jeyapriya, C.S.Kanimozhi Selvi **Publication:** IEEE, 2015	Frequent itemset mining, Naive bayes	92%
Title: Applying Supervised Opinion Mining Techniques On Online User Reviews [12] **Author:** Ion Smeureanu, Cristian Bucur **Publication:** Informatica Economică, 2012	Naive bayes, n-gram	80%

2.2 Dictionary Based Approach

In this approach first of all a small set of sentiment words which are known as seed words are collected manually with their known positive or negative orientations. Then this set is grown by searching their synonyms and antonyms in WordNet or another online dictionary. The new words are added to the existing seed list. Then next iteration is started. The iteration should be stopped when no new words are found. Manual inspection set is used at last to clean up the list.

In [18], Wordnet is used as dictionary Author uses mobile phone reviews from amazon website. It is input to the system. Polarity is calculated on the basis of majority of opinion words. Experimental results of 'AIRC Sentiment analyzer system' is compared with proposed system and proposed system provides better accuracy. In future, some enhancements in this technique will be carried out. It would deal with the sentences contain relative clauses like not only-but also and the sentences contain clauses neither-nor, either-or etc.

In paper [19] an Aspect based Opinion Mining system named as "Aspect based Sentiment Orientation System" is proposed which extracts the feature and opinions from sentences and determines whether the given sentences are positive, negative or neutral for each feature. Negation is also handled by the system. To determine the semantic orientation of the sentences a dictionary based technique of the unsupervised approach is adopted. To determine the opinion words and their synonyms and antonyms WordNet is used as a dictionary. All the features of the product on which reviews are given would be identified and the orientation of the sentence for each feature would be determined. The polarity of the given sentence is determined on the basis of the majority of opinion words. In the end the system will generate the feature wise summary of positive, negative and neutral sentences which will be easier for users to read, analyse and help them in taking the decision whether the product is to be purchased or not.

Table 2. Accuracy of Various Dictionary-based Algorithms

Title, Author, Publication	Method	Accuracy
Title: Mining Of Product Reviews At Aspect Level [19] **Author:** Richa Sharma, Shweta Nigam and Rekha Jain **Publication:** International Journal in Foundations of Computer Science & Technology (IJFCST), Vol.4, No.3, May 2014	Dictionary based unsupervised learning	74%
Title: Polarity Detection at Sentence Level [18] **Author:** Richa Sharma, Shweta Nigam, Rekha Jain **Publication:**International Journal of Computer Applications, Volume 86- No 11, 2014	Lexicon dictionary based approach	67%

3 ISSUES OF OPINION MINING

- A positive or negative sentiment world may have their opposite meaning in a particular domain so it is hard to predict by its keyword meaning [10].
- **Interrogative Sentence** An interrogative sentence may not have neither positive nor negative sentiment but the key word used in the opinion may be positive or negative[6].
- **Sarcastic Sentences** Few sentences in the form of jocks may violate the meaning of the whole sentences such kind of sentence need a power full attention toward the keywords and sentences. These funny sentences not only violet the sentence of a particular sentence but also destroy the value of the whole document[7].
- **Sentiment without sentiment words** sometimes sentiments does not use any sentiment words like good , better , best , worst ,bad etc. but the sentences may have its positive or negative feedback about the product , services and policies.
- **Conditional sentences** conditional sentences are also an issue in Sentiment mining conditional sentences is also creating the same problem like interrogative sentences[7].
- **Author and Reader understanding point (person to person varying)** Dollar price is increasing with respect to Indian rupee. This document have both the positive and negative meaning and its value is varying from person to person. This sentence has the positive sentiment for the Exporter while this same sentence has the negative sentiment for the importers [9].
- **Spam Reviews** Spam sentiments are those sentiments which are posted by the opposite or competitor organization for increasing their product value or their organization value among the users. Some politician may use the same spam review to just for their publicity.

4 APPLICATIONS OF OPINION MINING

- Business and e-commerce applications, such as product reviews and movie ratings [3]
- Opinions in the social and geopolitical context
- Predicting stock prices based on opinions that people have about the companies and resources [3]
- Determine areas of a product that need to be improved by summarizing product reviews to see what parts of the product are generally considered good or bad by users [4]
- Customer preference

5 CONCLUSION

Sentiment analysis has become very popular field of research. A lot has been researched in this field but still there are many issues as sentiment analysis processes text based unstructured data. Dictionary based approach takes less processing time than supervised learning approach but accuracy is not up to the mark. Supervised learning approach provides better accuracy. From this survey, it can be concluded that supervised techniques provide better accuracy compared to dictionary based approach.

In future, various opinion summarization algorithms should be applied to generate summary of all reviews provided by users.

6 REFERENCES

[1] Jiawei Han, Micheline Kamber and Jian Pei, “Data mining Concepts and Techniques”, Third Edition, Morgan Kaufmann Series in Data management Systems

[2] Charu C. Aggarwal, “Data Mining: The Textbook”, Springer, 2015

[3] Walaa Medhat, Ahmed Hassan, Hoda Korashy, "Sentiment Analysis Algorithms And Applications: A Survey", Ain Shams Engineering Journal, 2014

[4] Blessy Selvam1 , S.Abirami2, “A Survey On Opinion Mining Framework”, *International Journal of Advanced Research in Computer and Communication Engineering, 2013*

[5] G.Vinodhini, R.M.Chandrasekaran, "Sentiment Analysis And Opinion Mining: A Survey", International Journal of Advanced Research in Computer Science and Software Engineering, June 2012

[6] A.Jeyapriya, C.S.Kanimozhi Selvi,"Extracting Aspects And Mining Opinions In Product Reviews Using Supervised Learning Algorithm", IEEE, 2015

[7] Jeevanandam Jotheeswaran, Dr. Y. S. Kumaraswamy, "Opinion Mining Using Decision Tree Based Feature Selection Through Manhattan Hierarchical Cluster Measure", Journal of Theoretical and Applied Information Technology, 2013

[8] Kai Gao, Hua Xu, Jiushuo Wanga, "A Rule-Based Approach To Emotion Cause Detection For Chinese Micro-Blogs", ELSEVIER, 2015

[9] M.Ravichandran, G.Kulanthaivel, "Twitter Sentiment Mining (Tsm) Framework Based Learners Emotional State Classification And Visualization For E-Learning System", Journal of Theoretical and Applied Information Technology, 2014

[10] Chetashri Bhadane,Hardi Dalal, Heenal Doshi, "Sentiment Analysis: Measuring Opinions", Science Direct, 2015

[11] Jasakaran Kaur, Sheveta Vashisht, "Analysis And Indentifying Variation In Human Emotion Through Data Mining", Int.J.Computer Technology & Applications, 2012

[12] Ion Smeureanu, Cristian Bucur, "Applying Supervised Opinion Mining Techniques On Online User Reviews", Informatica Economică, 2012

[13] Weiyuan Li, Hua Xu, "Text-based emotion classification using emotion cause extraction", ELSEVIER, 2013

[14] Rizvaan Irfan, Christine K. King ,"A Survey on Text Mining in Social Networks", *The Knowledge Engineering Review, 2004*

[15] Diksha Sahni, Gaurav Aggarwal, "Recognizing Emotions and Sentiments in Text: A Survey ", International Journal of Advanced Research in Computer Science and Software Engineering , 2015

[16] Faiza Belbachir, B´en´edicte Le Grand, "Opinion Detection: Influence Factors", IEEE, 2015.

[17] Reshma Bhonde, Binita Bhagwat, Sayali Ingulkar, Apeksha Pande,"Sentiment Analysis Based on Dictionary Approach", International Journal of Emerging Engineering Research and Technology Volume 3, Issue 1,2015

[18] Richa Sharma, Shweta Nigam, Rekha Jain, "Polarity Detection at Sentence Level", International Journal of Computer Applications, Volume 86- No 11, 2014

[19] Richa Sharma,Shweta Nigam and Rekha Jain, "Mining Of Product Reviews At Aspect Level", International Journal in Foundations of Computer Science & Technology (IJFCST), Vol.4, No.3, May 2014.