A METHODOLOGICAL SURVEY ON SENTIMENT ANALYSIS TECHNIQUES AND THEIR APPLICATIONS IN OPINION MINING

MR. AVINAV PATHAK^{#1}, Dr. SAURABH SHARMA^{#2}, Mr. RAJESH PANDEY^{#3}

#1 Asst. Prof., CSE, Shobhit University, Meerut, 9456089798 and
avinav.pathak@shobhituniversity.ac.in

#2 Assoc. Dean, Vidya Knowledge Park, 7830202777 and saurabhsharma@vidya.edu.in

#3 Asst. Prof., CSE, Shobhit University, Meerut, 7417970699 and
rajesh@shobhituniversity.ac.in

ABSTRACT Opinion mining and sentiment analysis is a fast developing and growing research area. There are various e-commerce sites available on the net that provides choices to users to allow feedback regarding the specific product. This paper sharply focuses on survey of sentiment analysis techniques to check their effect on opinion mining. These feedbacks are significantly useful to both the people, who are willing to shop for that product and also the organizations. A correct technique for predicting sentiments may alter us, extract opinions from the web and predict customer's preferences. There are numerous algorithms available for opinion mining. Before applying any algorithmic rule for polarity detection, pre-processing on feedback is administered. From these pre-processed reviews opinion words and object on that opinion is generated are extracted and any opinion mining technique is applied to seek out the polarity of the review. Opinion mining has three levels of granularities: Document-level, Sentence level, and Aspect level. In this paper, numerous algorithms for sentiment analysis are studied and challenges and applications seem in this field are analyzed.

Key words: Sentiment Analysis, Opinion Mining, Web Content, Machine Learning, Aspect.

Corresponding Author: Dr. Saurabh Sharma

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 gather data regarding the positive and negative aspects of a specific topic. Finally, the positive and highly scored opinions obtained about a specific product are recommended to the user. To push marketing, massive firms and business individuals are making use of opinion mining [4].

Much analysis exists on sentiment analysis of user opinion data that primarily judge the polarities of user reviews. In these studies, sentiment analysis is usually conducted at one among the 3 levels: the document level, sentence level, or attribute level. About sentiment analysis, the literature survey done indicates 2 varieties of techniques together with machine learning and semantic orientation are vital [3]. These techniques are shown in figure one.

There are many challenges in Sentiment analysis. The primary may be an opinion word that's thought of to be positive in one scenario is also considered negative in another scenario. A second challenge is that people do not continuously express opinions in the same method. Most traditional text processing depends on the very fact that tiny variations between 2 items of text do not change the meaning significantly. In Sentiment analysis, however, "the image was great" is extremely different from "the image wasn't great". People will be contradictory in their statements [4]. Most reviews can have each positive and negative comment that is somewhat manageable by analyzing sentences one at a time.

Users express their opinions regarding products or services they consume in weblog posts, searching sites, or review sites. It's helpful for both the customers as well as for the producers to understand what the general public believes a particular product or service [6]. in the informal medium like twitter or blogs, the additional possible people are to combine different opinions within the same sentence that is straightforward for somebody's to know, however tougher for a computer to analyses. Typically even others have a problem understanding what somebody thought supported a brief piece of text as a result of it lacks context [5]

Fig 1 Different levels of opinion mining:

1. Document Level

In this approach, the whole document is considered as one entity and therefore the analysis approaches are applied on the full document. The result generated in the document level generally not acceptable [5] and [6].

2. Sentence Level

In the sentence level approaches, each sentence is taken into account as an entity and analysis approaches are applied to individual sentence then their results are summarized to produce the results of the document [5,6].

3. Aspect Level

Phrase-level opinion mining is additionally called aspect-based opinion mining. It performs fine-grained analysis and directly analyses at the opinion. The goal of this level of study is to find sentiments on aspects of items [5,6].

Explicit Aspects are explicitly mentioned as nouns or noun phrases in a sentence.

Implicit aspects are not explicitly mentioned in a sentence but are implied.

There are different types of algorithms to investigate sentiments. In this paper, all the techniques used for opinion mining is surveyed. Problems and applications of opinion mining are also mentioned. The rest of the paper is organized within the following approach. Section 2 offers a summary of the techniques used for opinion mining with its related work. Section 3 contains problems in opinion mining. Section 4 contains applications of opinion mining. Section 5 contains a conclusion.

2. RELATED WORK

2.1 Document Level

2.1.1 Supervised Learning Approach

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

2.1.1.1 Decision tree classifier

The decision tree classifier provides a hierarchical decomposition of the training data space in which a condition is used in the attribute value 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 that are used to classify them.

In [7] the film review features obtained from IMDB were extracted using the inverse document frequency and the importance of the word found. The principal components analysis and CART were used for the selection of characteristics according to the importance of the work concerning the whole document. The classification accuracy obtained by LVQ was 75%.

Exploring the emotional variation in adolescent age and the reasons behind these changes using data mining techniques are proposed in [11]. When classifying emotions and using the decision tree, different emotional variations are analyzed. The if-then rules are also generated from the decision tree. Typical value analysis is used to identify the variation of emotion in a child who has some type of disability.

2.1.1.2 Linear classifier

a. Support vector machine:

Text data is ideal for SVM classification due to the low nature of the text, in which few features are irrelevant, but tend to be correlated with each other and generally organized into linearly separable categories.

In [10], machine learning (SVM) combined with domain-specific lexicons is implemented for the classification of aspects and the polarity identification of the product review. SVM is trained to model the aspect classification and this trained SVM is used for the polarity classification by aspect. Experimental results indicate that the proposed techniques have achieved approximately 78% accuracy. The web-based data is applied to the emotional cause extraction subsystem and the complimentary feature selection method is combined, based on the result of these characteristics. In the training process, web publications with unknown

emotions are published in the SVM and SVR classification model and the result provides information on the type of emotion [13].

b. Neural network

The neural network consists of many neurons where the neuron is its basic unit. The inputs to the neurons are denoted by the vector on the line X_i , which is the frequency of words in the i^{th} document. There is a set of weights that are associated with each neuron used to calculate a function of its inputs. Based on inputs and weights, the output is generated.

2.1.1.3 Rule-based classifier

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

[8] Proposes a rules-based approach to the detection of emotion cause components for Chinese micro blogs. Present the emotion model and extract the corresponding cause components in fine-grained emotions. The emotional lexicon can be built manually and automatically from the corpus. Meanwhile, the proportions of the cause components can be calculated based on the multilingual characteristics based on Bayesian probability. The results of the experiment show the feasibility of the approach.

2.1.1.4 Probabilistic Classifier

a. Naïve Bayes'

The Naive Bayes classifier is the simplest and most used classifier. The Naive Bayes classification model calculates the subsequent probability of a class, based on the distribution of the words in the document. The model works with the extraction of BOW features that ignores the position of the word in the document. Use Bayes' theorem to predict the probability that a given set of characteristics belongs to a particular label.

The system proposed in [6] extracts aspects in customer reviews of products. Nouns and nominal phrases are extracted from each review sentence. The minimum support threshold is used to find all the frequent aspects of the review sentences given. The Naive Bayesian algorithm that uses a supervised approach based on counting terms is used to identify whether a sentence is a positive or negative opinion and also identifies the number of the sentence.

Article [12] presents a method of analyzing feelings about the review made by users to movies. The classification of the reviews in the positive and negative classes is based on a naive Bayes algorithm. As training data, we use a collection (pre-classified as positive and negative) of sentences taken from movie reviews. To improve the classification, we eliminated insignificant words and introduced them into word classification groups (n-grams). For $n = 2$ groups we achieved a substantial improvement in the classification.

b. Bayesian network

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

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.

Table 1 Accuracy of Various Supervised Algorithm

Title, Author, Publication	Author	Method	Accuracy
Objective Opinion Mining Using Decision Tree-Based Feature Selection Through Publication :Manhattan Hierarchical Cluster Measure [7] 2013	Jeevanandam Jotheeswaran, Dr. Y. S. Kumaraswamy	Naive Bayes	Precision: 0.50 Recall: 0.79
Objective : Sentiment Analysis: Measuring Opinions [10] Author Publication : Science Direct, 2015	Chetashri Bhadane,Hardi Dalal, Heenal Doshi	SVM	78%
Objective : Analysis And Identifying Variation In Human Emotion Through Data Mining [11] Author : Publication : Int.J Computer Technology & Applications, 2012	Jasakaran Kaur, Sheveta Vashisht	Decision Tree	NA
Objective : Twitter Sentiment Mining (Tsm) Framework Based Learners Emotional State Classification And Visualization For E-Learning System [9] Publication : Journal of Theoretical and Applied Information Technology, 2014	M.Ravichandran, G.Kulanthaivel	SVM, Maximum Entropy	95%, 95%
Title : A Rule-Based Approach To Emotion Cause Detection For Chinese Micro Blogs [8] Publication : ELSEVIER, 2015	Kai Gao, Hua Xu, Jiushuo Wanga	Association rule	75%
Title : Extracting Aspects And Mining Opinions In Product Reviews Using Supervised Learning Algorithm [6] Publication : IEEE, 2015	A.Jeyapriya, C.S.Kanimozhi Selvi	Frequent itemset mining, Naive Bayes	92%
Title : Applying Supervised Opinion Mining Techniques On Online User Reviews [12] Publication : Informatica Economică, 2012	Ion Smeureanu, Cristian Bucur	Naive Bayes, n-gram	80%

2.1.2 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 the next iteration is started. The iteration should be stopped when no new words are found. The manual inspection set is used at last to clean up the list.

In [18], Wordnet is used as a dictionary The author uses mobile phone reviews from the Amazon website. It is input to the system. Polarity is calculated based on the majority of opinion words. Experimental results of the AIRC Sentiment analyzer system" are compared

with the proposed system and the proposed system provides better accuracy. In the 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 “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 like 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 based on 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, analyze and help them in deciding 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%

- **Sentence Level**

At this level, the task is to determine each sentence in the document as positive or negative opinions. The sentiment analysis at the prayer level has classified polarity. This level is close to the level of the document, but here it is achieved with each sentence [20]. However, there may be complex sentences in the text that make the level of the sentence not useful. There are two phases in the feeling of the level sentence that is performed in each sentence: first, each sentence classified, as subjective or objective, and the second is the polarity of the subjective sentence that is concluded.

- **Aspect Level**

Online opinions based on aspects expressed by users on social networking sites have become a popular source of information for consumers regarding their purchasing decisions, as well as for companies seeking opinions about their products. Therefore, it is important to develop opinion mining applications based on aspects with emphasis on extracting and classifying opinions based on aspects expressed by users about the products in a given review. Previous studies have used a limited set of heuristic patterns for the extraction of aspects with

algorithms for the classification of feelings related to both supervised (based on annotated data sets) and unsupervised (based on lexical resources)[21]. New methods are designed to increase the user utility for studying the aspects of various entities on web 3.0 and other platforms.

• Evaluation of Sentiment Classification

In general, the performance of the classification of feelings is evaluated by using four indexes: accuracy, precision, recall and F1 score. The common way to calculate these indices is based on the confusion matrix shown in Table 1.

	Possible Positives	Possible Negatives
Actual Positive Schema	Number of True Positive Datasets (TPD)	Number of False Negative Datasets(FND)
Actual Negative Schema	Number of False Positive Datasets (FPD)	Number of True Negative Datasets(TND)

These indexes have been defined by the following mathematical relations:

$$\text{Accuracy} = \frac{TPD + TND}{TPD + TND + FPD + FND} \dots\dots\dots(i)$$

$$\text{Precision} = \frac{TPD}{TPD + FPD} \dots\dots\dots(ii)$$

$$\text{Withdrawal} = \frac{TPD}{TPD + FND}$$

$$\text{F1 Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Accuracy is the portion of all true predicted instances against all predicted instances. An accuracy of 100% means that the predicted instances are the same as the actual instances. Precision is the portion of true positive predicted instances against all positively predicted instances. Withdrawal is the portion of true positive predicted instances against all real positive instances. F1 score is a harmonic average of accuracy and recovery.

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 either positive or negative sentiment but the keyword 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 needs powerful 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 do 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.

- d) **Conditional sentences:** are also an issue in Sentiment mining conditional sentences and are also creating the same problem as interrogative sentences [7].
- e) **Author and Reader understanding point (person to person varying):** Dollar price is increasing concerning Indian rupee. This document has both positive and negative meaning and its value is varying from person to person. This sentence has a positive sentiment for the Exporter while this same sentence has a negative sentiment for the importers [9].
- f) **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 politicians may use the same spam review just for their publicity.

4. APPLICATIONS OF OPINION MINING

- a) Commercial and e-commerce applications, such as product reviews and film ratings [3]
- b) Opinions in the social and geopolitical context.
- c) Predict share prices based on the opinions people have about companies and resources [3]
- d) Determine the areas of a product that should be improved by summarizing product reviews to see which parts of the product are generally considered good or bad by users [4]
- e) Customer preference

5. CONCLUSION

Sentiment analysis has become a very popular research field. There has been a lot of research in this field, but there are still many issues because sentiment analysis processes structured data based on text. A dictionary-based approach requires less processing time than a supervised learning approach, but the accuracy is not equal. The supervised learning process provides greater accuracy. From this survey, it can be concluded that the monitored methods provide greater accuracy compared to the dictionary-based approach. In this document, we have introduced problems of classification of feelings at different levels, that is, document level, prayer level, word level, and aspect level. Additionally, some techniques are introduced that are accustomed to solving these problems. In the future, more research is required to enhance the performance metrics of the methods and techniques presented in this document. In the future, many feedback summary algorithms should be applied to generate a summary of all the revisions offered by users.

6. REFERENCES

- [1] J Han, M. Kamber, and J. 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] W. Medhat, A. Hassan, H. Korashy, "Sentiment Analysis Algorithms And Applications: A Survey", *Ain Shams Engineering Journal*, 2014
- [4] B. 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] J. 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 Wang, 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] C.Bhadane, H. Dalal, Heenal Doshi, Sentiment Analysis: Measuring Opinions, *Science Direct*, 2015
- [11] J. Kaur, S.Vashisht, Analysis And Identifying Variation In Human Emotion Through Data Mining, *Int.J.Computer Technology & Applications*, 2012
- [12] I.Smeureanu, C. Bucur, Applying Supervised Opinion Mining Techniques On Online User Reviews, *Informatica Economică*, 2012
- [13] W.Li, Hua Xu, Text-based emotion classification using emotion cause extraction, *ELSEVIER*, 2013
- [14] R. Irfan, Christine K. King, A Survey on Text Mining in Social Networks, *The Knowledge Engineering Review*, 2004
- [15] D. Sahni, G.Aggarwal, Recognizing Emotions and Sentiments in Text: A Survey, *International Journal of Advanced Research in Computer Science and Software Engineering*, 2015
- [16] F. Belbachir, B'en'edict Le Grand, "Opinion Detection: Influence Factors", *IEEE*, 2015.
- [17] R. Bhonde, B. 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] R. Sharma, S. Nigam, Rekha Jain, "Polarity Detection at Sentence Level", *International Journal of Computer Applications*, Volume 86- No 11, 2014
- [19] R.Sharma, S.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.
- [20] H. Yu and V. Hatzivassiloglou. "Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. *In Proceedings of the 2003 conference on Empirical methods in natural language processing*, EMNLP '03 pages 129-136.
- [21] Asghar, M.Z., Ahmad, S., Qasim, M., Zahra, S.R., Kundi, F.M.: SentiHealth: creating health-related sentiment lexicon using the hybrid approach. *Springer Plus*. 5(1):1139 (2016)