International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

# A Survey on Sentiment Analysis Technique in Web Opinion Mining

**S. Veeramani[1], S. Karuppusamy[2]**

[1]PG Scholar (Computer Science and Engineering), Nandha Engineering College, Erode, India

[2]Assistant Professor, Department of Computer Science, Nandha Engineering College, Erode, India

**Abstract:** ***This survey covers techniques and approaches that promise to directly enable opinion-oriented information-seeking systems. Our focus is on methods that seek to address the new challenges raised by sentiment-aware applications, as compared to those that are already present in more traditional fact-based analysis. We include material on summarization of evaluative text and on broader issues regarding privacy, manipulation, and economic impact that the development of opinion-oriented information-access services gives rise to. To facilitate future work, a discussion of available resources, benchmark datasets, and evaluation campaigns is also provided.***

**Keywords:** Opinion mining, sentiment analysis, sentiment lexicon, feature extraction, sentiment classification

## 1. Introduction

A Sentiment analysis and opinion mining are subfields of machine learning. They are very important in the current scenario because, lots of user opinionated texts are available in the web now. This is a hard problem to be solved because natural language is highly unstructured in nature. The interpretation of the meaning of a particular sentence by a machine is tiresome. But the usefulness of the sentiment analysis is increasing day by day. Machines must be made reliable and efficient in its ability to interpret and understand human emotions and feelings. Sentiment analysis and opinion mining are approaches to implement the same. The sentiment analysis problem can be solved to a satisfactory level by manual training. But a fully automated system for sentiment analysis which needs no manual intervention has not been introduced yet. This is mainly because of the challenges in this field. This paper aims at a literature survey on the problem of sentiment analysis and opinion mining. Many relevant studies have emerged in this field and this paper is a peep into some of them.

Opinion Mining (OM), a promising discipline, defined as combination of information retrieval and computational linguistic techniques deals with the opinions expressed in a document [12]. The field aims at solving the problems related to opinions about products, politics in newsgroup posts, review sites, etc [13]. There are different techniques for summarizing customer reviews like Data Mining, Information Retrieval, Text Classification and Text Summarization [13].

Opinion Mining or Sentiment Analysis is the field to extract the opinionated text datasets and summarize in understandable form for end user [15]. Opinion mining is to extract the positive, negative or neutral opinion summary from unstructured data.World Wide Web users asked the opinions of his family and friends to purchase the product. In the same way when organizations needed to take the decision about their products they had to conduct the surveys to the focused groups or they had to hire the external consultants [5, 13]. Web 2.0 [14], facilitate the customers to take decision to purchase the product by reviewing the posted comments. Customers can post reviews on web communities, blogs, discussion forums, twitters, product's web site these comments are called user generated contents [12]. Web2.0 is playing a vital role in data extracting source in opinion mining. It facilitates users to know about the product from other customer's reviews that have already used it instead of asking friends and families. Companies, instead of conducting surveys and hiring the external consultants to know about the consumers opinions, extract opinionated text from product web site [13, 15]. An automated opinion summarization model is needed to perform these tasks. It is the sub-discipline of web content mining, involves Natural Language Processing and opinion extraction task to find out the polarity of any product consumers feedback [1].

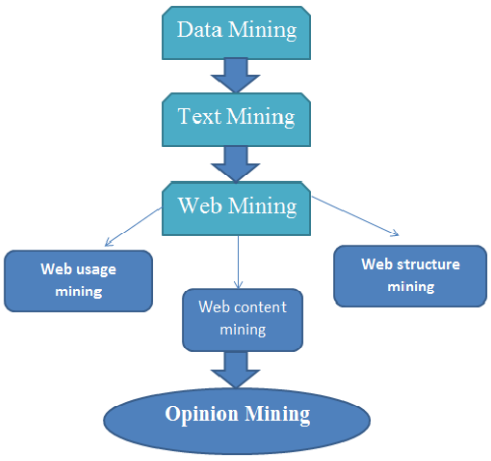


**Figure 1:** Opinion Mining Overview

This paper is organized as follows: section 2 covers components of opinion mining. Section 3 is about the levels of sentiment analysis. Section 4 describes techniques in opinion mining. Section 5 defines the opinion mining and summarization. Section 6 covers research areas in opinion mining.

## 2. Components of Opinions

1. Opinion Holder: Opinion holder is the person or organization that expresses the opinion
2. Opinion Object: It is a feature about which the opinion holder is expressing his opinion.
3. Opinion Orientation: Determine whether the opinion about an object is positive, negative or neutral.

## 3. Levels of Sentiment Analysis

### 3.1. Document Level Sentiment Analysis

The basic information unit is a single document of opinionated text. In this document level classification, a single review about a single topic is considered. But in the case of forums or blogs, comparative sentences may appear. Customers may compare one product with another that has similar characteristics and hence document level analysis not desirable in forums and blogs. The challenge in the document level classification is that the entire sentence in a document may not be relevant in expressing opinion about an entity. Therefore subjectivity/objectivity classification is very important in this type of classification. The irrelevant sentences must be eliminated from the processing works. Document level sentiment classification executed on the overall sentiments expressed by authors Documents classified according to the sentiments instead of topic. It is to summarize the whole document as positive or negative polarity about any object (mobile, car, movie, and politician).

### 3.2. Sentence level sentiment analysis

In the sentence level sentiment analysis, the polarity of each sentence is calculated. The same document level classification methods can be applied to the sentence level classification problem. Objective and subjective sentences must be found out. The subjective sentences contain opinion words which help in determining the sentiment about the entity. After which the polarity classification is done into positive and negative classes. In case of simple sentences, a single sentence bears a single opinion about an entity. But there will be complex sentences also in the opinionated text. In such cases, sentence level sentiment classification is not desirable. Knowing that a sentence is positive or negative is of lesser use than knowing the polarity of a particular feature of a product. The advantage of sentence level analysis lies in the subjectivity/objectivity classification. The traditional algorithms can be used for the training processes.

### 3.3. Phrase level sentiment analysis

The phrase level sentiment classification is a much more pinpointed approach to opinion mining. The phrases that contain opinion words are found out and a phrase level classification is done. This can be advantageous or disadvantageous. In some cases, the exact opinion about an entity can be correctly extracted. But in some other cases, where contextual polarity also matters, the result may not be fully accurate. Negation of words can occur locally. In such cases, this level of sentiment analysis suffices. But if there are sentences with negating words which are far apart from the opinion words, phrase level analysis is not desirable. Also long range dependencies are not considered here. The words that appear very near to each other are considered to be in a phrase.

## 4. Techniques Used in Opinion Mining

Database contains the important hidden information used for decision making. Different databases like relational, object oriented, transactional and spatial databases consist on the complex dataset. Major data mining techniques used to extract the knowledge and information are:

generalization, classification, clustering, association rule mining, data visualization, neural networks, fuzzy logic, Bayesian networks, genetic algorithm, decision tree, multi agent systems, CRISP-DM model, churn prediction, Case Based Reasoning and many more.

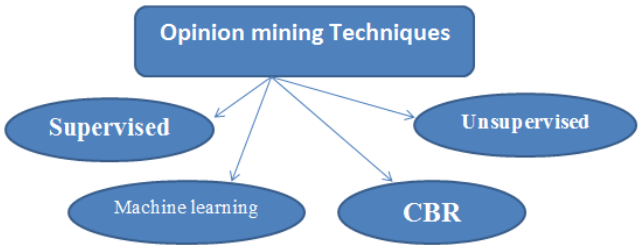


**Figure 2:** Opinion Mining Technique

rapidgrowth in databases has created the need to develop such technologies to extract the suggest of knowledge and information intelligently. Data mining techniques are most suitable for this purpose, these techniques directly refers Artificial Intelligence. Major data rule mining, data visualization, neural networks, fuzzy logic, Bayesian networks, genetic algorithm, mining techniques used to extract the knowledge and information are: generalization, classification, clustering, association decision tree, multi agent systems, churn prediction, Case Based Reasoning, techniques, association and many more.

### 4.1. Supervised Machine Learning

Classification is most frequently used and popular data mining technique. Classification used to predict the possible outcome from given data set on the basis of defined set of attributes and a given predictive attributes. The given dataset is called training dataset consist on independent variables (dataset related properties) and a dependent attribute (predicted attribute). A training dataset created model test on test corpora contains the same attributes but no predicted attribute. Accuracy of model checked that how accurate it is to make prediction. Classification is a supervised learning used to find the relationship among attributes.

### 4.2. Unsupervised Learning

In contrast of supervised learning, unsupervised learning has no explicit targeted output associated with input. Class label for any instance is unknown so unsupervised learning is about to learn by observation instead of learn by example. Clustering is a technique used in unsupervised learning. The process of gathering objects of similar characteristics into a

group is called clustering. Objects in one cluster are dissimilar to the objects in other clusters.

### 4.3. Case Based Reasoning

Case based reasoning is an emerging Artificial Intelligence supervised technique used to find the solution of a new problem on the basis of past similar problems. CBR is a powerful tool of computer reasoning and solve the problems (cases) in such a way which is closest to real time scenario. It is a recent problem solving technique in which knowledge is personified as past cases in library and it does not depend on classical rules. The past problem's solutions are stored in CBR repository called Knowledge base or Case base. Instead of solving the new problem by "first principal" reasoning, CBR use the knowledge base to reuse the The solution of past similar problem if needed to the In case base repository as a new solution instance in CBR cycle consists of four R's. Nowadays it is the most emerging technique used in opinion mining systems. Statistical methods are combined with knowledge extracting techniques in to enhance case searching, browsing and Reuse it for the problem solving methods semantic analysis of a sentence in natural language that can be easily used and manipulated in a textual data mining process. This sentence analysis uses and depends on several types of knowledge that are: a lexicon, a case base and hierarchy of index. In this methodology a case based reasoning model is adopted that is based on the classification rules and course of similarity for the assurance of the compliance.

## 5. Opinion Mining and Summarization Process

Opinion Mining also called sentiment analysis is a process of finding user's opinion towards a topic. Opinion mining concludes whether user's view is positive, negative, or neutral about product, topic, event etc. Opinion mining involves analyzing user's opinion, attitude, and emotion towards particular topic. This consists of first categories text into subjective and objective information, and then finding polarity in subjective text. Opinion mining can be performed word, sentence or document level. Opinion mining and summarization process involve three main steps, first is Opinion Retrieval, Opinion Classification and Opinion Summarization.

Summarization of opinions is a major part in opinion mining process. Summary of reviews provided should be based on features or subtopics that are mentioned in reviews. Therefore, feature extraction [4] and opinion summarization are key issues. Many researchers worked on summarization product reviews [2]. The opinion summarization process mainly involve following two approaches. One is Feature based summarization another one is Term Frequency based summarization.

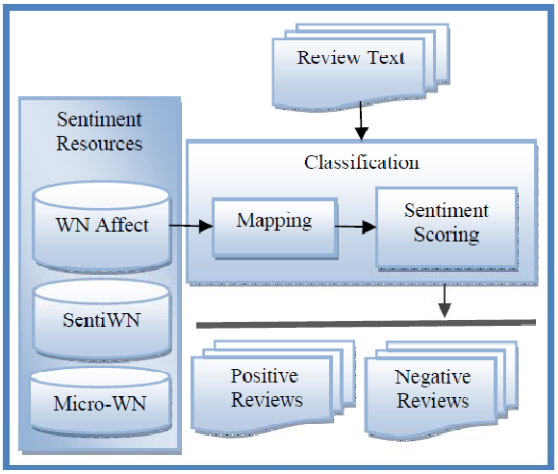


**Figure 3:** Opinion Mining Summarization

### 5.1Opinion Retrieval

Opinion retrieval is a process of collecting reviews text from review websites. Different review websites contain reviews for products, movies, hotels, news etc. Information retrieval techniques such as web crawler can be applied to collect review text data from many sources and store them in database. This step involves retrieval of reviews, micro-blogs, comments etc of user. We should only consider the data which contain subjective data but not the objective data. Reviews are retrieved by query based information retrieval techniques.

### 5.2. Opinion Classification

Primary step in sentiment analysis is classification of review text. Given a review document D = {d1..d1} and predefined categories set *C = {positive, negative},* sentiment classification is to classify each *di* in *D*, with expressed in *C*. There are many approaches for sentiment classification in opinion text. Machine learning and lexicon based approach is more popular.

#### 5.2.1. Machine learning approach for opinion classification:

The machine learning approach uses supervised learning method for classification of review text. The first step is to train a classifier using sample of reviews with its class (positive/negative). Then the built model of trained classifier is used to predict category of new text reviews. Popular machine learning classifiers for text categorization are Support Vector Machines (SVM) and Naive Bayes(NB).

#### 5.2.2. Lexicon based approach for opinion classification:

The lexicon Approach predicts sentiment of review text using databases which contain word polarity values e.g. SentiWordNet [10]. Review text is classified by calculating and averaging polarity score of individual words in sentences. Many factors such as word position, word relationships, negation handling should be considered while sentiment classification using lexicon based approach.

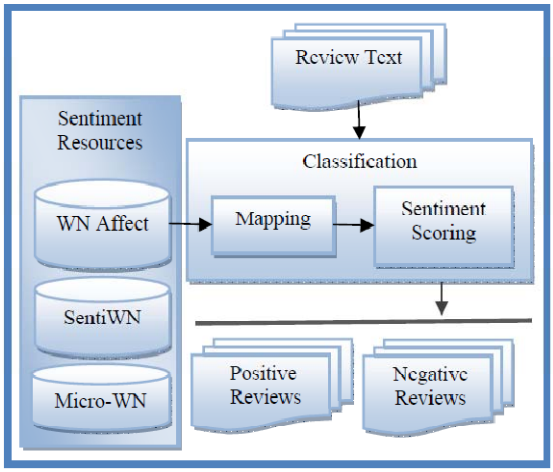


**Figure 4:** Lexicon based Opinion Classification

### 5.2.3 Opinion Summarization

Summarization of opinions is a major part in opinion mining process. Summary of reviews provided should be based on features or subtopics that are mentioned in reviews. Therefore, feature extraction and opinion summarization are key issues. Many researchers worked on summarization product reviews. The opinion summarization process mainly involve following two approaches.

**1) Feature based summarization:**

This type summarization involve finding of frequent terms (features) that are appearing in many reviews. The summary is presented by selecting sentences that contain particular feature information. Features present in review text can be identified using Latent Semantic Analysis (LSA) method. For a short summary of product reviews, product features and opinion words associated.

**2) Term Frequency based summarization:**

Term frequency is count of term occurrences in a document. If a term has higher frequency it means that term is more import for summary presentation. In many product reviews certain product features appear frequently and associated with user opinions about it. In this method sentences are scored by term frequency. The summary is presented by selecting sentences that are relevant and which contain highest frequency terms. Opinion Summarization process is shown in Fig.5 It shows review text is preprocessed which involve sentence segmentation and tokenization of sentence in terms. After calculating term frequency of each term, each sentence score and relevance is calculated. As per the compression rate highest scoring and relevant sentences are presented in summary.

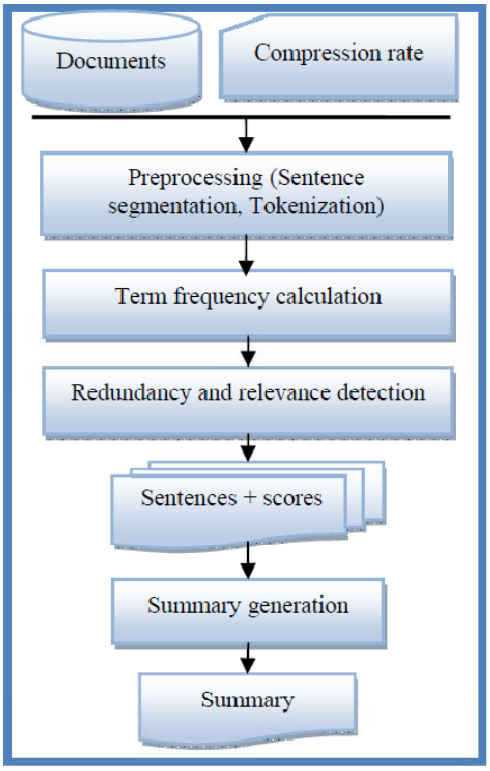


**Figure 5:** Opinion summarization

## 6. Research Areas in Opinion Mining

Current research is focusing on

1) Customer Reviews for Individual Product Feature Based Ranking.
2) Overall positive and negative polarity at paragraph level.
3) Ranking of best paragraph or sentence based on best feature and their polarity involved.
4) Continuous Improvement of the algorithms for opinion detection.
5) Decrease the human effort needed to analyze contents.
6) Semantic analysis through lexicon/corpus of words with known sentiment for sentiment classification.
7) Identification of highly rated experts.

### 6.1 Future Research: long term and short term issues

**Short-term:**

- Enhanced discoverability of content through Linked Data
- Visual representation
- Audiovisual opinion mining
- Real-time opinion mining
- Machine learning algorithms
- SNA applied to opinion and expertise
- Bipolar assessment of opinions
- Multilingual reference corpora
- Comment and opinion recommendation algorithm
- Cross-platform opinion mining
- Collaborative sharing of annotating/labeling resources

**Long-term:**

- Autonomous machine learning and artificial intelligence
- Usable, peer-to-peer opinion mining tools for citizens
- Non-bipolar assessment of opinion
- Automatic irony detection

Sentiment detection has a wide variety of applications in information systems, including classifying reviews, summarizing review and other real time applications. There are likely to be many other applications that are not discussed. It is found that sentiment classifiers are severely dependent on domains or topics. From the above work it is evident that neither classification model consistently outperforms the other, different types of features have distinct distributions. It is also found that different types of features and classification algorithms are combined in an efficient way in order to overcome their individual drawbacks and benefit from each other's merits, and finally enhance the sentiment classification.

## 7. Conclusion

Due to web and social network, large amount of data are generated on Internet every day. This web data can be mined and useful knowledge information can be fetched through opinion mining process. This paper discussed different opinion classification and summarization approaches, and their outcomes. This study shows that machine learning approach works well for sentiment analysis of data in particular domain such as movie, product, hotel etc., while lexicon based approach is suitable for short text in micro-blogs, tweets, and comments data on web.

## References

[1] N. Au, R. Law, and D. Buhalis. The impact of culture on ecomplaints: Evidence from the chinese consumers in hospitality organization. In U. Gretzel, R. Law, and M. Fuchs, editors, Information and Communi- cation Technologies in Tourism 2010, pages 285–296. Springer Verlag Wien, 2010.

[2] C. Chen, F. Ibekwe-SanJuan, E. SanJuan, and C. Weaver. Visual analysis of conflicting opinions. In IEEE Symposium n Visual Analytics Science And Technology, pages 35 – 42, 2006.

[3] Ziqiong Zhang, Qiang Ye, Zili Zhang, Yijun Li, "Sentiment classification of Internet restaurant reviews written in Cantonese", ExpertSystem with applications,2011.

[4] Padmaja.S and SameenFatima ,"Opinion Mining and Sentiment Analysis –An Assessment of Peoples‟ Belief: A Survey", International Journal of Ad hoc, Sensor & Ubiquitous Computing (IJASUC) Vol.4, No.1, February 2013

[5] KaiquanXu , Stephen Shaoyi Liao , Jiexun Li, Yuxia Song, "Miningcomparative opinions from customer reviews for Competitive Intelligence",Decision Support Systems 50 ,743–754, (2011).

[6] G. Jaganadh 2012. Opinion mining and Sentiment analysis CSI communication.

[7] Yingcai Wu, Furu Wei, Shixia Liu, Norman Au, Weiwei Cui, Hong Zhou, and HuaminQu, *Member, IEEE* "Opinion Seer: Interactive Visualization of Hotel Customer Feedback,"IEEE transactionson visualization and computer graphics,vol.16,no.6,november/december 2010.

[8] Bing Liu. Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012.

[9] C.L. Liu, W.H. Hsaio, C.H. Lee, G.C. Lu and E. Jou, "Movie Rating and Review Summarization in Mobile Environment", *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews*, Vol. 42, No. 3, pp. 397-407, 2012.

[10] B. Pang, L. Lee, and S. Vaithyanathan,"Thumbs up?:Sentiment classification using machine learning techniques,"inProc.ACL-02Conf.Empirical Methods Natural Lang. Process., 2002, pp. 79–86.

[11] Chien-Liang Liu, Wen-Hoar Hsaio, Chia-Hoang Lee, Gen-Chi Lu, and Emery Jou, "Movie Rating and Review Summarization in Mobile Environment",IEEE VOL. 42, NO. 3, MAY 2012.

[12] Pang, B., L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classi_cation using machine learning techniques. In Proceedings ofConference on Empirical Methods in Natural Language Processing (EMNLP-2002), 2002.

[13] Ku, L.-W., Liang, Y.-T., & Chen, H, "Opinion extraction, summarization and tracking in news and blog corpora". In AAAI-CAAW‟06.

[14] Melville, WojciechGryc, "Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification", KDD09, June 28–July 1, 2009, Paris, France.Copyright 2009 ACM 978-1-60558-495-9/09/06.

[15] Titov, I., McDonald, R.: A Joint Model of Text and Aspect Ratings for Sentiment Summarization. In: Proceedings of ACL-2008: HLT, pp. 308–316 (2008).

[16] Nilesh M. Shelke, Shriniwas Deshpande, PhD. and Vilas Thakre, PhD., Survey of Techniques for Opinion Mining, International Journal of Computer Applications (0975 – 8887) Volume 57– No.13, November 2012.

[17] Xiaohui Yu, Member, IEEE, Yang Liu, Member, IEEE, Jimmy Xiangji Huang, Member, IEEE, and Aijun An, Member, IEEE, "Mining Online Reviews for Predicting Sales Performance: A Case Study in the Movie Domain," IEEE Transactions on Knowledge and Data Engineering, Vol. 24, NO. 4,APRIL 2012.

[18] Christopher C. Yang and TobunDorbin Ng, *Member, IEEE,* "Analyzing and Visualizing Web Opinion Development and Social Interactions With Density-BasedClustering,"IEEE Transactions on Systems, man, and cybernetics—part a: systems and humans, vol. 41, no. 6, novemBER 2011.

[19] AinurYessenalina, YisongYueandClaireCardie,Multi-level Structured ModelsforDocument- level Sentiment Classification, Proceedings of the 2010 Conference on Empirical Methods in Natural LanguageProcessing,pages1046–1056,MIT,Massachusetts, USA,9-11October 2010.AssociationforComputationalLinguistics.

[20] Hanhoon Kang, SeongJoonYoo, Dongil Han," Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews". Expert Systems with Applications 39 (2012) 6000–6010.

[21] Qiang Ye, Ziqiong Zhang, Rob Law, "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches", Expert Systems with Applications 36 (2009) 6527–6535.

## Author Profile

**S. Veeramani** received the B.E. degree in Computer science and Engineering from Nandha engineering college in 2012. He is currently doing his M.E Computer science and Engineering in Nandha engineering college, Erode, India