International Journal of Information and Computation Technology.
ISSN 0974-2239 Volume 3, Number 8 (2013), pp. 737-742

http://www. irphouse.com /ijict.htm

# Supervised Opinion Mining Techniques: A Survey

**Richa Sharma[1], Shweta Nigam[2] and Rekha Jain[3]**

*[1]Department of Computer Science, Banasthali University*
*C-62, Sarojini Marg, C-Scheme, Jaipur, India.*

**Abstract**

Opinions play a very important role for human beings. Whenever a decision has been taken, opinions of others are also considered. Nowadays millions of web-users express their opinions about many topics through blogs, chats and social networks. Businesses and organizations always want to find consumer or public opinions about their products and services. In e-commerce and e-tourism, it is very important to automatically analyze the huge amount of social information available on the Web; therefore, it has become important to develop methods of automatically classifying them. Opinion Mining or Sentiment Analysis is the mining of attitudes, opinions, and emotions automatically from text, speech, and database sources through Natural Language Processing (NLP). This survey paper gives an overview of the supervised techniques that classify the opinions as positive or negative.

**Keywords**: Sentiment Analysis, Opinion Mining, Machine learning.

## 1. Introduction

Human beings are always keen to know what other peoples think. Whenever a decision has to be taken, the opinions of friends and relatives are always considered. But nowadays where the Internet is used by everyone, there is no need to consult with anyone because there are lots of reviews available online which helps us to know whether the product is good or bad. Here, the Opinion Mining plays an important role. Opinion Mining can be defined as a sub-discipline of computational linguistics that focuses on extracting opinion of persons from the web. It is a *Natural Language Processing (NLP)* and *Information Extraction (IE)* task that aims to obtain feelings of

writer expressed in positive or negative comments by analyzing a large number of documents [Shelke et al., 2012].

Opinion Mining combines the techniques of computational linguistics and information retrieval and is concerned with the opinions expressed rather than topics in the text. Opinions can be expressed on anything.

Three main components of Opinion Mining are:

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.

For example "This mobile has an excellent *voice quality*". In this review, Opinion Holder is the user who has written this review. Opinion object here is the *voice quality* of the mobile phone and the opinion word is "excellent" which is positively orientated. Determination of semantic orientation is a task of concluding whether a sentence or document has either positive or negative orientation. [Pang et al., 2002] and [Turney, 2002] reported early works attempting this task. This task can be decomposed into two approaches: the unsupervised approach [Turney,2002] and the supervised approach .This survey paper mainly focuses on the supervised approach of Sentiment Classification. The remainder of this paper is organized into the following sections: Section 2 explains various supervised techniques in detail. Section 3, include in references discussed about existing research work .The last section concludes the study.

## 2. Sentiment Analysis Approaches

Sentiment Analysis or Opinion Mining aims to determine the attitude of a speaker or a writer with respect to some topic. There are two types of techniques mainly used in opinion mining:

- Machine Learning
- Semantic Orientation.

The machine learning approach belongs to supervised classification approach. This approach is more accurate because each of the classifiers is trained on a collection of representative data known as corpus. Thus, it is called "supervised learning".

In a machine (supervised) learning based classification, two types of documents are required: training set and test set. A training set is used to learn the classifier and a test set is used to test the performance of the automatic classifier. Large numbers of machine learning techniques are available which classifies the opinions. Machine learning techniques like Naïve Bayes, Maximum Entropy (ME) and Support Vector Machines (SVM) have achieved great success in text categorization.

**2.1 Naive Bayes Classification**

A Naive Bayes Classifier is a simple probabilistic classifier based on Bayes' theorem and is particularly suited when the dimensionality of the inputs are high. Naïve Bayes classification is an approach to text classification that assigns the class c to a given document *d.*

$$c* = \arg\max_c P(c|d) \quad (1)$$

The *Naive Bayes* (NB) classifier uses the Bayes rule given in eq(2)

$$P(c|d) = \frac{P(c)P(d|c)}{P(d)} \quad (2)$$

Where P(c|d) is the probability of instance d being in class c , P(d|c) is the probability of generating instance d given class c , P(c) is the probability of occurrence of class c and P(d) is the probability of instance d occurring. To estimate the term $P(d \mid c)$, Naive Bayes decomposes it by assuming the fi's are conditionally independent given in eq(3)

$$P_{NB}(c|d) := \frac{P(c)\left(\Pi_{i=1}^{m} P(f_i|c)^{n_i(d)}\right)}{P(d)} \quad (3)$$

Where $P_{NB}(c|d)$ is the probability of instance d being in class c, m is the no of features and $f_i$ is the feature vector. Naive Bayes is optimal for certain problem classes with highly dependent features [ Domingos et al., 1997].

**2.2 Maximum Entropy**

Maximum Entropy (ME) classification is yet another technique, which has proven effective in a number of natural language processing applications [Berger et al., 1996]. Its estimate of $P(c \mid d)$ takes the exponential form as given in eq (4).

$$P_{ME}(c|d) := \frac{1}{Z(d)} \exp\left(\sum_i \lambda_{i,c}\ F_{i,c}\ (d, c)\right) \quad (4)$$

Where $P_{ME}(c|d)$ is the probability of instance d being in class c, $\lambda$ is Lagrange multipliers, $Z(d)$ is a normalization function. *Fi,c* is a *feature/class function* for feature *fi* and class *c*, as in eq(5).

$$F_{i,c}(d, c') := = \begin{cases} 1, n_i(d) > 0 \ and \ c' = c \\ 0 \ otherwise \end{cases} \quad (5)$$

Maximum Entropy provides better performance when conditional independence assumptions are not met.

**2.3 Support Vector Machines**

Support vector machines (SVMs) have been shown to be highly effective at traditional text categorization. They are *large-margin*, rather than probabilistic, classifiers. The

basic idea behind the training procedure is to find a maximum margin hyper plane, represented by vector $\underset{\omega}{\rightarrow}$, that not only separates the document vectors in one class from those in the other, but for which the separation, or *margin*, is as large as possible. This corresponds to a constrained optimization problem; letting $c_j$ {1, −1} (corresponding to positive and negative) be the correct class of document $d_j$, the solution can be written as in eq (6).

$$\underset{\omega}{\rightarrow} = \sum_j \alpha_j c_j \underset{d_j}{\rightarrow}, \alpha_j \geq 0 \quad (6)$$

Where $\underset{\omega}{\rightarrow}$ is a vector, $c_j$ is a class and $d_j$ is a. document. Those $d_j$ for which $\alpha_j$ are greater than zero are called support vectors. Classification of test instances consists simply of determining which side of $\underset{\omega}{\rightarrow}$ hyper plane they fall on.

## 3. Related Work

Most researchers focus on assigning sentiments to documents. Some of the existing researches in sentiment analysis using supervised techniques are as follows: Songbo Tan et al., has performed sentiment analysis on Chinese documents. He investigated four feature selection methods (MI,IG, CHI and DF) and five learning methods (winnow classifier, K-nearest neighbor, centroid classifier, Naive Bayes and SVM) on a Chinese sentiment corpus. From the results he concludes that, IG performs the best for sentimental terms selection and SVM exhibits the best performance for sentiment classification. Rudy Prabowo et al., has combined rule-based classification, machine learning and supervised learning method. For each sample set, they carried out 10-fold cross validation and for every fold, the associated samples were divided into training and a test set. For each test sample, a hybrid classification is carried out.

Ziqiong Zhang et al., proposed a method which utilizes completely prior-knowledge-free supervised machine learning method .They performed sentiment analysis on written Cantonese. Their method has proved that the chosen machine learning model could be able to draw its own conclusion from the distribution of lexical elements in a piece of Cantonese review. Long-Sheng Chen et al., proposed a neural network based approach, which combines the advantages of the machine learning techniques and the information retrieval techniques. Lina Zhou et al.,[7] investigated movie review mining using machine learning and semantic orientation. Supervised classification and text classification techniques are used in the proposed machine learning approach to classify the movie review. A corpus is formed to represent the data in the documents and all the classifiers are trained using this corpus. Their experimental results showed that the supervised approach is more efficient.

Bo Pang et al , used machine learning techniques for sentiment analysis. The experimental setup consists of movie-review corpus with randomly selected 700 positive sentiment and 700 negative sentiment reviews. Learning methods Naïve Bayes, maximum entropy classification and support vector machines were employed.

Experiments demonstrated that the machine learning techniques are better than human produced baseline for sentiment analysis on movie review data. Alekh Agarwal et al.,proposed a machine learning method that incorporates linguistic knowledge gathered through synonymy graphs, for opinion classification. This approach shows the degree of influence among relationships of documents have on their sentiment analysis. Ahmed Abbasi et al., proposed sentiment analysis method that utilized the function of stylistic and syntactic features to evaluate the sentiment in English and Arabic content. The Entropy weighted Genetic Algorithm is incorporated to enhance the performance of the classifier and achieve the true assessment of the key features. Experiments were conducted using movie review data set and the results demonstrated that the proposed techniques are efficient.

## 4. Conclusion

Opinion Mining plays a very important role in making a decision about product or services. Opinion Mining has large application areas like Education in which opinion can be used to evaluate academics based on opinions expressed by students. Shopping, where websites like amazon.com allow customers to express their opinions on their websites. Entertainment where the people's can easily see the reviews of their favorite movies and daily soaps online. Marketing, Companies can now make savings on marketing expenses by requesting for reviews on their websites. Now there is no need to conduct surveys as companies can now have all the data they need online. In this survey several machine learning techniques have been discussed and the related work has been done by using these techniques. But, still there are some of the challenges that still to be resolved like entity identification, negation handling, complexity in handling sentence or document etc. Researchers have been carried out to overcome these challenges.

## References

[1] Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra (1996), A maximum entropy approach to natural language processin, Computational Linguistics, *Journal Computational Linguistics*, **22**,*1*,pp 39–71.

[2] Alekh Agarwal & Pushpak Bhattacharyya(2005), Sentiment Analysis: A new approach for effective use of linguistic knowledge and exploiting similarities in a set of documents to be classified, *In Proceedings of the International Conference on Natural Language Processing (ICON)*.

[3] Ahmed Abbasi, Hsinchun Chen and Arab Salem (2008),Sentiment Analysis in Multiple Languages:Feature Selection for Opinion Classification in Web Forums,*ACM Transactions on Information Systems*,**26**,*3*.

[4] BoPang and Lillian Lee (2008), Opinion Mining and Sentiment Analysis,*Foundations and Trends In Information retrieval*,**2**,*1-2*,pp 1-135.

[5] Lina Zhou,Pimwadee Chaovalit (2005),Movie review mining:a comparison between Supervised and Unsupervised Classification Approaches,*Proceedings of the 38th Hawaii International Conference on system sciences.*

[6] Long Sheng Chen,Cheng-Hsiang Liu,Hui-Ju Chiu(2011),A neural network based approach for sentiment classification in the blogosphere,*Journal of Informetrics*,**5,***2,*pp 313-322.

[7] Nilesh M.Shelke, Shriniwas Deshpande,Vilas Thakre(2012),Survey of techniques for opinion mining,*International Journal of Computer Applications,* **57,***13,pp 0975-8887.*

[8] Pedro Domingos and Michael J. Pazzani,(1997),On the optimality of the simple Bayesian classifier under zero-one loss machine learning,**29,**pp 103-130.

[9] Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques,*Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*,pp 79-86.

[10] Qiang Ye, Ziqiong Zhang, Rob Law (2009),Sentiment classification of online reviews to travel destinations by supervised machine learning approaches, *Expert systems with applications,* **36,** *3,pp 6527-6535.*

[11] Rudy Prabowo, Mike Thelwall (2009), Sentiment analysis: A combined approach,*Journal of Informetrics*,**3**,*2*,pp 143-157.

[12] Songbo Tan , Jin Zhang (2008),An empirical study of sentiment analysis for Chinese documents , *Expert Systems with applications*,**34**,*4*,pp 2622–2629.

[13] Turney, P (2002), Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, *In* proceedings *of 40th Annual Meeting of the ACL,* pp. 417-424.