ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 2, February 2017

A Survey on Sentiment Analysis and Opinion Mining

[1] Rasika Wankhede [2] Prof. A. N. Thakare
[1] M. Tech Computer Science and Engineering [2] Assistant Professor,
Department of Computer Engineering, Bapurao Deshmukh College of Engineering, Sewagram.

Abstract: **-- As whole world is changing rapidly and using the current technologies such as internet, has become an essential need for everyone. Now a day, large number of users or suggestions on any topic is present on web. Opinions may contain the reviews on product, hotels, or the reviews on movies, which helps other users in their decision making. Opinion mining is what public thinks about a particular topic, as it is an open source platform, every individual has rights to give their opinions. Public opinion plays an important role in various sectors. As opinions are present in the form of positive and negative polarity this means that it is present in the form of good and bad sentiments. Sentiments are emotions, a specific view or judgment on certain topic. Sentimental analysis is used for classifying polarity for the given text document. In this paper various algorithms for sentiment analysis are studied and challenges and applications appear in this field are discussed.**

Keywords:— **Opinion mining, Polarity, Sentiment, Sentimental Analysis.**

I. INTRODUCTION

Opinion mining is one of the new concepts of data mining. In opinion mining opinions of various peoples are mine from the reviews for particular topic. Opinion mining system which is able to collect texts from web based conventional media and social media. The opinions are calculated in the form of polarity such as positive opinions or negative opinions. There are various sectors where opinion mining can be used such as shopping sites, movies, hotels, tourism etc. It is domain of natural language processing and text analytics. Opinion mining is described as the processing of unstructured data and text data to categorize it into some result like positive, negative and neutral so that we can predict the product. It produces the text qualities from textual sources.Opinion mining plays an important role in the process of information retrieval or knowledge discovery from huge amount of data [1]. There are three levels of opinion mining.

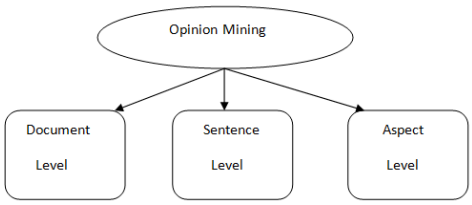

Fig. 1: Levels of opinion mining

1) Document Level

In document level analysis approach is applied on the whole document and the whole document is considered as single entity.

2) Sentence Level

In sentence level analysis approach is applied on every individual sentence and each sentence is considered as an entity.

3) Aspect Level

Aspect level opinion mining is also known as phrase-level opinion mining. The main goal of this level is to discover sentiments on aspects of items.

There are various applications of opinion mining such as:

- Opinions in social and geopolitical context.
- Application in Government Intelligence knowing consumer demeanors and patterns.
- Business and e-commerce application, such as product reviews and movie ratings.
- Stock price are predicted based on opinions that people have about companies and resources.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 2, February 2017

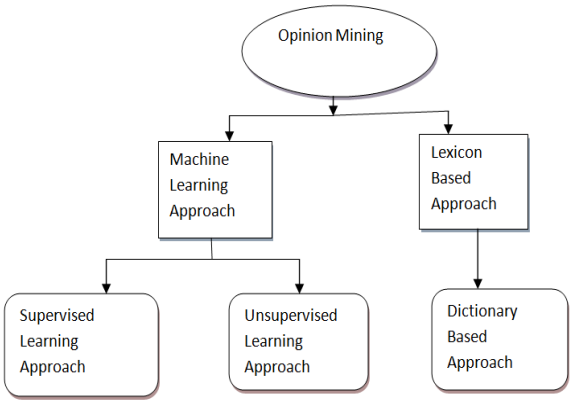

Fig. 2: Techniques of Opinion Mining.

Machine learning approach is sub divided into two categories such as supervised learning approach and unsupervised learning approach and Lexicon based approach contain dictionary based approach.

A. Supervised Learning Approach

In this method there are two sets of documents which are training set and test set. The training set is used by classifier for learning about the document and the test set is used for the validation purpose. Types of supervised learning methods are:

1) Decision tree classifier

Decision tree classifier offers the method for hierarchical decomposition of training data space, where the condition of attribute value is used for dividing the data. The condition is absence or presence of one or more words [3]. The division process is performed recursively until the leaf nodes contain minimum records for process of classification.

2) Linear classifier

a) Support Vector Machine

In machine learning SVM are supervised learning models with associated learning algorithm that analyze data used for classification and regression analysis [3]. The clustering algorithms which provide an important support vector machine is called support vector clustering and it is used in industrial application either when data are not labeled or when only some data are labeled as a preprocessing for a classification pass.

b) Neural Network

Neural network consist of many neurons where neuron is its basic unit. The main goal of neural network is to solve problem in the same way as that the human brain can solve, although several neural network are abstract. The set of weights are present which is associated with each neuron in order to compute function of its inputs. Depending on input the output weight is obtained.

3) Rule based classifier

In rule based classifier, the set of rules are applied to data space. The left hand side shows the condition on feature set which is represented in disjunctive normal form where as the class label is present on the right hand side.

4) Probabilistic classifier

a) Naïve Bayes

In machine learning, naïve bayes classifiers are a family of simple probabilistic classifier based on applying Baye's theorem with strong independence assumption between the features [3]. Naïve Bayes are highly scalable, requiring a number of variables (features/predictors) in learning problem. Naïve Bayes is a simple technique for constructing classifiers models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite sets.

b) Bayesian Network

There are two assumptions of the Naïve Bayes classifier first is Naïve Bayes classifier is the independent of features and other is to assume that all the features are fully dependent. Bayesian Network is considered as a complete model for the variables and their relationships. It is not frequently used because its computation complexity is very expensive.

c) Maximum Entropy

The maximum entropy classifier uses encoding technique for converting labeled feature sets to vectors. The encoded vector is used for calculating weights of each feature.

B. Dictionary Based Approach

In this approach the seed words which are small set of sentiment words are collected with their known

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 2, February 2017

positive or negative orientations. The new words are added to the list and then next iteration is performed the iteration process stops when new words are found. Manually the seed list is cleaned by using inspection set. Sentiment analysis is used for classifying the polarity such as positive polarity, negative polarity or neutral for the given text of the document. Sentiment analysis is sub domain of opinion mining where the analysis is focused on the extraction of emotions and opinions towards particular topic [1].

Sentiment analysis is not directly applied on any data, firstly reviews are extracted from web sites and then pre-processing is performed to obtain more precise result [2]. However, to classify the polarity of opinion is not an easy task as it includes many challenges, the first challenges is the opinion word that is considered to be positive in one situation may be considered negative in another situation for example: Opinions are expressed by people in different ways and sometime one opinion may contain combination of opinions but it become difficult for computer system to parse. In normal text processing if small difference is present between two sentences then both are considered as same sentences but in sentimental analysis, a small difference may change the polarity of sentence. For example, "Dangal movie is good" is totally different from " Dangal movie is not good". Another challenge is that people never express their opinions in the same way every person has different opinions regarding certain topics.

Organization of this paper provides the following details: Section II discusses the related work done in this domain. Section III explains the proposed work in this paper in depth. Section IV gives the conclusion for the proposed work.

II. RELATED WORK

A large number of works have been carried out previously on opinion mining and sentimental analysis. Nagamma P et al. [1] proposed different data mining techniques to online movie review data and also predicted the box office collection for the movie. Classification accuracy for pretending was improved substantially by clustering. The online movie review data collected from web sites, the box office collection and the success or failure of the movie is predicted based on the reviews.

Pang et al. [2] applied machine learning, methods for classifying the online movies reviews, collected from the internet movie database, to positive or negative opinions, by obtaining the list of 14 affective keywords which are then applied in a straightforward way to form the baseline for classification. Sentiment analysis classifies the opinions into positive and negative categories. R.M. Chandrasekaram et al. [3] stated 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 merits, and finally enhance the sentiment classification performance. Wang et.al [4] proposed supervised learning method have been widely employed and proven effective in sentiment classification. They normally depend on a large amount of labeled data. To overcome this problem various semi-supervised learning methods are proposed to effectively utilize a small scale of labeled data along with large amount of unlabelled data.

Turney et al. [5] proposed document level sentiment classification there are two kinds of approaches, term counting approaches usually involve deriving a sentiment measure by calculating the total number of negative and positive terms. L. Lee et al. [6] sentiment analysis attempt to automatically identify and recognize opinions and emotion in the text in the form of positive polarity, negative polarity. Domingos P.et al. [7] concluded that Naïve Bayes provides efficient results in case of certain problems where features are highly dependent. Whereas, Naïve Bayes has basic assumption that features must be independent to each other but Naïve Bayes gives better results. Dr. Y.S. Kumaraswamy et al. [8] Movies review features obtained from IMDb was extracted using inverse document frequency and the importance of word found. Kai Gao et al. [9] proposed a rule-based approach to emotion cause component detection for Chinese micro-blogs. It extracts the corresponding cause components in fine grained emotions. A. Jayapriya et al. [10] extracts aspects in

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 2, February 2017

product customer reviews. Naïve Bayesian classification algorithm using supervised term counting based approach is used to identify the sences is positive or negative opinion. Ion Smeureanu et al. [11] proposed method of sentiment analysis, on the review made by users on movies. To improve classification insignificant words were removed and group of words in classification was introduced. Kennedy and Inkpen et al. [12] evaluated a negation model which is fairly identical to model proposed by Polanyi in document level polarity classification. Gang Li et al. [13] developed an approach based on the k-means clustering algorithm. Rui xia et al. [14] proposed an ensemble technique which combines the output of several base classification models to form an integrated output. This approach is comparative study of effectiveness of ensemble technique for sentiment classification. Songho tan et al. [15] stated the idea to estimate the probabilities of categories given a text document by using joint probabilities of words and categories. The similarity score of each nearest neighbor document to the text document is used as the weight of the classes of neighbor document. Dave, Lawrence and Pennock [16] also use machine learning methods to explore sentiment classification. However, they select top words according to their generated points instead of using all the words. Tony Mullen and Nigel Collier et al. [17] use SVM to analyze sentiment orientation of words as well as topic-oriented and artist oriented information. Pak and Paroubek et al. [18] develop a sentiment classifier for twitter data using words-bag method relying on features from twitter corpus, which shows the application of sentiment analysis in social network. Qingliang Miao et al. [19] proposed a sentiment mining and retrieval system. The sentiments are mine from the huge amount of information using lexical resources. M. Ravichandran et al. [20] use SVM to analyze sentiment in the form of emotions from Twitter. The classification and visualization is carried out for E- Learning systems. Melville et al. [21] proposed the method for sentiment analysis of blogs by combining lexical knowledge with text classification using Bayesian classifier.

III. PROPOSED WORK

This section gives the detailed description of the steps followed for the movie data set mining for sentiment classification and for obtaining the accurate result. The opinions are collected from the movie reviews on websites. These opinions are present in the form of text, various process are performed on collected data in the form of opinions and the output obtained in the form of positive opinion, negative opinion or neutral opinions.

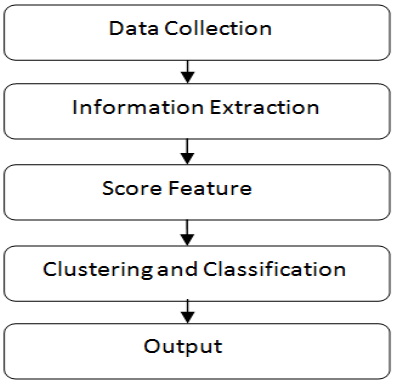

Fig. 3: The steps involved in sentiment classification.

A. Data Collection

Opinions regarding various topics are present on various websites. The data collection components collect data from specific data sources. Opinions present on various web sites are useful for users to take correct decision. There are many fields where opinion mining is used such as news, blogs, twitter, movie, hotel, tourism etc. The raw data is collected from various data sources and given as input for information extraction process.

B. Information Extraction

In the process of information extraction the useful information is extracted from huge dataset. The useful information is in the form of sentiments that are represented in the form of positive polarity or negative polarity. The extracted information is provided as input for further processing.

C. Score Features

The score features are selected to decide the opinions polarity. Opinions are collected from various data sources, features are selected to decide whether opinion belongs to positive polarity, negative polarity or neutral. There is certain opinion which does not show any polarity (positive/negative) is called as neutral opinions.

D. Clustering and Classification

The clustering is performed after the selection of score features; various clusters are formed by

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 2, February 2017

collecting the relevant information and then forming the clusters. The different clusters are formed for positive features, negative features as well as for neutral features. The classification process is performed to classify the sentiments. Sentiment classification mainly consists of two tasks, such as sentiment intensity assignment and sentiment polarity assignment. Sentiment intensity assignment deals with analyzing, whether positive or negative sentiments are mild or strong. Sentiment polarity assignment deals with analyzing whether a text has a positive, negative or neutral semantic orientation. The final output provides the exact opinion based on polarity which can be represented in graphical form.

TABLE I: Accuracy of various algorithms for sentiment analysis.

Sr.No	***AUTHOR NAME, TITLE***	***Method***	***Accuracy***
1.	Author: Kennedy and Inkpen (2006). Title: Sentiment classification using movie reviews using contextual valence shifters.	SVM	86.2%
2.	Author: Godbole et al. (2007) Title: Large scale sentimental analysis for news and blogs.	Lexical approach	82.7-95.7%
3.	Author: Songho tan(2008) Title: An empirical study of sentiment analysis for Chinese document.	Centroid classification and SVM	90% S
4.	Author: Qingliang Miao(2009). Title: "AMAZING: A sentiment mining and retrieval system.	Lexical resource	87.6%
5.	Author: Melville et al. (2009) Title: Sentiment analysis of blogs by combining lexical knowledge with text classification.	Bayesian classification	91.21%
6.	Author: Gang Li et al. (2010) Title: Twitter sentiment classification using distant supervision.	K-means clustering	78%
7.	Author: Rui Xia et al. (2011) Title: Ensemble of feature sets and classification algorithm for sentiment classification.	Naïve bayes, maximum entropy, SVM	NB-85.8% ME-85.4% SVM-86.4%
8.	Author: Ion smeureanu (2012) Title: Applying supervised opinion mining technique on online user reviews.	Naïve Bayes	80%
9.	Author: Dr. Y.S. Kumaraswamy et al. (2013) Title: Opinion mining using decision tree based on feature selection through Manhattan clustering measure.	Naïve Bayes	79%
10.	Author:M. Ravichandran et al. (2014) Title: Twitter sentiment mining framework based learners emotional state classification and visualization for E-Learning system.	SVM	95%
11.	Author: Jayapriya et al.(2015) Title: Extraction Aspects and mining opinions in product reviews using supervised learning algorithm.	Naïve Bayes	92%

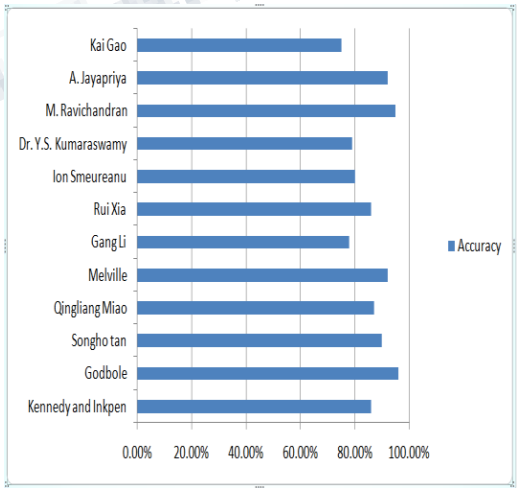

Fig. 4: Accuracy achieved in sentimental analysis.

IV. CONCLUSION

Opinion mining helps the user to obtain the exact reviews related to the topic. From last decade,

many researchers worked on this area of opinion mining. Mining techniques are proposed for sentimental analysis. Many challenges have been solved by new techniques; different tools are also available for sentimental analysis. A lot has been researched in this field but still there are many issues as sentimental analysis processes text based unstructured data. From this survey it can be concluded that supervised technique provide better accuracy for sentimental analysis.

V. ACKNOWLEDGEMENT

This work is a part of the postgraduate level project work and I represent my sincere gratitude to Professor A. N. Thakare, BDCOE, Sewagram, for his constant guidance throughout the work and support. And all the staff member of Computer Science and Engineering department, Bapurao Deshmukh College of Engineering for providing me excellent atmosphere for Dissertation work.

REFERENCES

[1] P.Nagamma, Pruthvi H.R, Nisha K.K, Carlos Soares," An ImprovedSentiment Analysis of Online Movie Reviews", IEEE 2015, International conference on Computer and Inforamation Technology.

[2] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?:sentiment classification using machine learning techniques," in Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002, pp. 79–86.

[3] R.M. Chandrasekaram, G.Vinodhini "Sentimental Analysis and Opinion Mining- A Survey International Journal of Advanced Research in Computer Science and Software Engineering.

[4] Li, S., Wang, Z., Zhou, G., & Lee, S.Y.M. (2011), 'Semi-supervised learning for imbalanced sentiment classification', In Proceedings of international joint conference on artificial intelligence, pp. 1826–1831.

[5] Turney, Peter, and Michael L. Littman. "Unsupervised learning of semantic orientation from a hundred-billion-word corpus." (2002).

[6] Lee L. Measures of distributional similarity. 1999: Proceedings of ACL.25-32.

[7] Domingos P. and Pazzani, M.1997. On the optimality of the simple Bayesian classifier under zero-one loss, Machine Learning, vol. 29, no. 2-3, pp. 103–130.

[8] 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.

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

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

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

[12] Kennedy and D. Inkpen, "Sentiment classification of movie reviews using contextual valence shifters," Computational Intelligence, vol. 22, pp. 110–125,2006.

[13] Gang Li , Fei Liu , "A Clustering-based Approach on Sentiment Analysis" ,2010, 978-1-4244-6793-8/10,2010 IEEE.

[14] Rui Xia , Chengqing Zong, Shoushan Li, "Ensemble of feature sets and classification algorithms for sentiment classification", Information Sciences 181 (2011) 1138–1152.

[15] Songbo Tan , Jin Zhang, "An empirical study of sentiment analysis for chinese documents ", Expert Systems with Applications 34 (2008) 2622–2629.

[16] K. Dave, S. Lawerence & D. Pennock, "Mining the Peanut Gallery-Opinion Extraction and Semantic Classification of Product Reviews", Proceedings of the 12th International World Wide Web Conference, pp. 519-528, 2003.

[17] Tony Mullen, Nigel Collier, "Sentiment analysis using support vector machine with diverse information sources", National Institute of Informatics Hitotsubashi 2-1-2, Chiyoda-ku, Tokyo.

[18] Alexander Pak, Patrick Paroubek ,"Twitter as a corpus for sentiment analysis and opinion mining", Universite de Paris-Sud, laboratoire LIMSI-CNRS, Batiment 508, F-91405 Orsay Cedex, France.

[19] Qingliang Miao, Qiudan Li, Ruwei Dai , "AMAZING: A sentiment mining and retrieval system", Expert Systems with Applications 36 (2009) 7192–7198.

[20] 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.

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