ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

A Survey on Sentiment Analysis and Opinion Mining

Dudhat Ankitkumar M[1], Prof. R. R. Badre[2], Prof. Mayura Kinikar[3]

PG Scholar, Dept. of CE, MIT Academy of Engineering, Pune, India[1]

Associate Professor, Dept. of CE, MIT Academy of Engineering, Pune, India[2]

Assistant Professor, Dept. of CE, MIT Academy of Engineering, Pune, India[3]

ABSTRACT: Sentiment Analysis (SA), which is also called opinion mining, is the field of study which analyzes people's opinions, sentiments, evaluations, appraisals, attributes and emotions towards entities such as products services, organizations, individuals, issues, events, topics. SA is machine learning approach in which machine analyzes and classifies the human's sentiments, emotions, and opinions about some topic which are expressed in the form of either text or speech. SA aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. SA is an ongoing research field. This survey paper contains overview of the recent updates in this field. Many recent algorithms and various SA applications are explained briefly in this paper. The main target of this paper is to give detailed explanation of SA techniques and the related fields. The purpose of this paper is the illustration of the recent trend of research in the sentiment analysis and its related areas.

KEYWORDS: Sentiment Analysis, Opinion Mining, Machine learning, Sentiment lexicon.

I. INTRODUCTION

Sentiment Analysis (SA) and Opinion Mining (OM) 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. SA or OM is the computational study of people's opinions, attitudes and emotions towards entity. The entity can represent individuals, events or topics. These topics are most likely to be covered by reviews. The two expressions SA or OM are interchangeable.

There are many challenges in sentiment analysis. The first is that an 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 express opinions in a same way. Most reviews will have both positive and negative comments, which somewhat manageable by analyzing sentences one at a time. However in more informal medium like twitter or blogs, the more likely people are to combine different opinions in the same sentence which is easy for human to understand, but more difficult for a computer to parse. Sometimes even other people have difficulty to understand what someone's thought based on a short piece of text because it lacks context.

Main fields of research in sentiment analysis are Subjectivity Detection, Sentiment Prediction. Aspect based Sentiment Summarization, Text summarization for Opinions, Contractive viewpoint Summarization, Product Feature Extraction, Detecting opinion spam.

Subjectivity Detection is a task of determining whether text is opinionated or not. Sentiment prediction is about predicting the polarity of text whether it is positive or negative. Aspect based Sentiment summarization provides sentiment summary in the form of star ratings or scores of features of product. Text summarization generates a few sentences that summarize the reviews of a product. Contrastive viewpoint summarization puts an emphasis on contradicting opinions. Product feature Extraction is a task that extracts the product features from its review. Detecting opinion spam is concern with identifying fake or bogus opinion from reviews.

Sentiment analysis can be done at Document level, Sentence level, and Aspect or Feature level. In Document level the whole document is classified either into positive or negative class. Sentence level sentiment classification classifies

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

sentence into positive, negative or neutral class. Aspect or Feature level sentiment classification concerns with identifying and extracting product feature from the source data.

The sentiment analysis problem can be solved to a satisfactory level by manual training. But a fully automated system for sentiment analysis which needs to manual intervention has not been introduced yet. This is the main challenge of this field.

There are two main approaches for sentiment analysis: machine learning based and lexicon based. Machine learning based approach uses classification technique to classify text. Lexicon based method uses sentiment dictionary with opinion words and match them with data to determine polarity. They assigns sentiment scores to the opinion words describing how Positive, Negative and Objective the words contained in the dictionary are.

The objective of this paper is to discover the concept of Sentiment Analysis in the field of Natural Language Processing and present a comparative analysis of its techniques in this field.

Other sections in this paper show different levels of SA and different classification techniques respectively.

II. DIFFERENT LEVELS OF SENTIMENT ANALYSIS

A. *Document Level Sentiment Analysis*

Basically information is a single document of opinionated text. A single review about a single topic is considered in this document level classification. But comparative sentences may appear in the case of forums or blogs. In forums and blogs sometimes document level analysis is not desirable when customer may compare one product with another that has similar characteristics. The challenge in the document level classification is that all the sentence in an entire document may not be relevant in expressing the opinion about an entity. So subjectivity/objectivity classification is very much important in this type of classification.

For document level classification both supervised and unsupervised learning methods can be used. Any supervised learning algorithm like naïve Bayesian, Support Vector Machine, can be used to train the system. The unsupervised learning can be done by extracting the opinion words inside a document. Thus the document level sentiment classification has its own advantages and disadvantages. Advantage is that we can get an overall polarity of opinion text about a particular entity from a document. Disadvantage is that the different emotions about different features of an entity could not be extracted separately.

B. *Sentence Level Sentiment Analysis*

The same document level classification methods can be applied to sentence level classification problem. In the sentence level sentiment analysis, the polarity of each sentence is calculated. Subjective and objective sentences must be found out. The subjective sentences contain opinion words which help to determine the sentiment about an entity. After which the polarity classification is done into positive and negative classes. In case of simple sentences, a single sentence contains a single opinion about an entity. But in case of complex sentence in the opinionated text sentence level sentiment classification is not done. Getting the information that 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.

C. *Phrase Level Sentiment Analysis*

This classification is much more pinpointed approach to opinion mining. The phrases that contain opinion are found out and a phrase level classification is done. In some cases, the exact opinion about an entity can be correctly extracted.
But in some cases negation of words can occur locally. In these cases, this level of sentiment analysis suffices. The words that appear very near to each other are considered to be in a phrase.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

III. SENTIMENT CLASSIFICATION TECHNIQUES

Sentiment classification techniques can be roughly divided into machine learning approach, lexicon based approach and hybrid approach [5]. The *Machine Learning Approach (ML)* applies the famous ML algorithms and uses linguistic features. The *Lexicon-based Approach* relies on a sentiment lexicon, a collection of known and precompiled sentiment terms. It is divided into dictionary-based approach and corpus-based approach which use statistical or semantic methods to find sentiment polarity. The *Hybrid Approach* combines both approaches and is very common with sentiment lexicons playing a key role in the majority of methods.

The text classification methods using *Machine learning* approach can be divided into supervised and unsupervised learning methods. The supervised methods use a large number of labelled training documents. The unsupervised methods are used when it is difficult to find these labelled training documents.

The lexicon-based approach depends on finding the opinion lexicon which is used to analyze the text. The dictionary-based approach which depends on finding opinion seed words, and then searches the dictionary of their synonyms and antonyms. The corpus-based approach starts with a seed list of opinion words, and then finds other opinion words in a large corpus to help in finding opinion words with context specific orientations. This could be done by using statistical or semantic methods.

A. *Machine Learning*

Machine learning approach relies on the famous Machine Learning algorithms to solve the sentiment analysis as a regular text classification problem that makes use of syntactic and/or linguistic features.

Text Classification Problem Definition: We have a set of training records $D = \{X_1, X_2, \ldots\ldots, X_n\}$ where each record is labeled to a class. The hard classification problem is when one label is assigned to an instance. The soft classification problem is when a probabilistic value of labels is assigned to an instance.

I.A.1 *Supervised Learning*

The supervised learning methods depend on the existence of labelled training documents. There are many supervised classifiers in literature. Next subsection, we present it with brief details about some of the most frequently used classifier in Sentiment Analysis.

I.A.1.1 *Probabilistic Classifiers*

Probabilistic classifiers use mixture models for classification. The mixture model assumes that each class is a component of the mixture. These kind of classifiers are also called generative classifiers because each mixture component is a generative model that provides the probability of sampling a particular term for that component. Three of the most famous probabilistic classifiers are discussed in next subsections.

I.A.1.1.1 *Naive Bayes Classifier(NB)*

Naïve Byes classifier is the simplest and most commonly used classifier. Naïve Bayes classification model computes the posterior probability of a class based on the distribution of the words in the document. It uses Bayes Theorem to predict the probability that a given feature set belongs to a particular label.

$$P(label|features) = \frac{P(label)*P(features|label)}{P(features)} \quad (1)$$

P(label) is the prior probability of a label or the likelihood that a random feature set label. *P(features|label)* is the prior probability that a given feature set is being classified as a label. *P(feature)* is the prior probability that a given feature set is occurred. Given the Naïve assumption which states that all features are independent, the equation could be rewritten as follows:

$$P(label|features) = \frac{P(label)*P(f1|label)*.....*P(fn|label)}{P(features)} \quad (2)$$

An improved NB classifier was proposed by Kang and Yoo [2] to solve the problem of tendency for the positive classification accuracy to appear up to approximately 10% higher than the negative classification accuracy. They showed that using this algorithm with restaurant reviews narrowed the gap between the positive accuracy and the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

negative accuracy compared to NB and SVM. The accuracy is improved in recall and precision compared to both NB and SVM.

I.A.1.1.2 *Bayesian Network(NB)*
The main assumption of the NB classifier is the independence of 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 [1].

BN was used by Hernandez and Rodriguez [3] to consider a real-world problem in which the attitude of author is characterized by three different but related target variables. They propose the use of multi-dimensional Bayesian network classifiers. It joined the different target variables in the same classification task in order to exploit the potential relationships between them. They showed that their semi-supervised multi-dimensional approach out performs the most common SA approaches, and that their classifier is the best solution in a semi-supervised framework because it matches the actual underlying domain structure.

I.A.1.2 *Linear Classifiers*
Given $\bar{X} = \{x_1 \dots . x_n\}$ is the normalized document word frequency, vector $\bar{A} = \{a_1 \dots . a_n\}$ is a vector of linear coefficients with the same dimensionality as the feature space, and b is a scalar; the output of the linear predictor is defined as $p = \bar{A}.\bar{X} + b$, which is the output of the *linear classifier*. The prediction p is a separating hyper plane between different classes. There are many kinds of linear classifiers; among them is *Support Vector Machines (SVM)* [4,6] which is form of classifiers that attempt to determine good linear separators between classes. Two of most famous linear classifiers are discussed in the following subsections.

I.A.1.2.1 *Support Vector Machines Classifiers (SVM).*
The main principle of SVM is to determine linear separators in the search space which can best separate the different classes. In Fig. 1 there are 2 classes *x, o* and there are 3 hyperplanes A, B and C. Hyperplane is provide the best separation between the classes, because the normal distance of any of the data points is the largest, so it represents the maximum margin of separation.

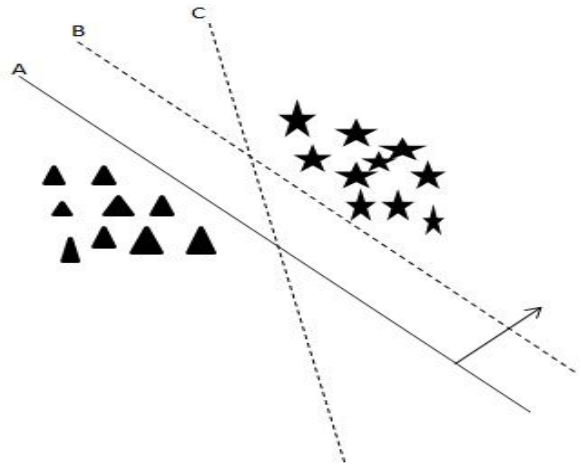

Fig-1: Using support vector machine on a classification problem.

The 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 [8]. SVM can construct a nonlinear decision surface in the original feature space by mapping the data instances non-linearly with a hyper plane [7].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

As a discriminative model, SM uses $g(x) = w^T \emptyset(x) + b$ as the discriminant function, where w is the weights vector, b is the bias, and $\emptyset(\bullet)$ denotes nonlinear mapping from input space to high dimensional feature space. The parameters w and b are learned automatically on the training dataset following the principle of the maximized margin by

$$\min \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi i \quad (3)$$
$$\text{s.t.} \begin{cases} yig(xi) \geq 1 - \xi i \\ \xi i \geq 0, i = 1, \ldots\ldots, N, \end{cases}$$

Where ξi denotes the slack variables and C is the penalty coefficient.
The training sample $(\overline{Xi}, yi)$ is called a support vector when satisfying the Lagrange multiplier αi>0. By introducing kernel function discriminant function can be represented as

$$g(x) = \sum_{i=1}^{\bar{N}} \alpha iyiK(xi, x). \quad (4)$$

Due to the dimension of feature space is quite large in text classification tasks, the classification problem is always linearly separable [21], therefore linear kernel is commonly used.

SVMs are used in many applications, among these applications are classifying reviews according to their quality. Chen and Tseng [9] have used two multiclass SVM-based approaches: One-versus-All SVM and Single-Machine Multiclass SVM to categorize reviews. They proposed a method for evaluating the quality of information in product reviews considering it as a classification problem. They also adopted an information quality (IQ) frame work to find information oriented feature set. They worked on digital cameras and MP3 reviews. Their results showed that their method can accurately classify reviews in terms of their quality.

SVMs were used by Li and Li [10] as a sentiment polarity classifier. Unlike the binary classification problem, they argued that opinion subjectivity and expresser credibility should also be taken into consideration. They proved that their mechanism can effectively discover market intelligence (MI) for supporting decision-makers by establishing a monitoring system to track external opinions on different aspects of a business in real time.

I.A.2 *Weakly, Semi and Unsupervised Learning*
The main goal of the classification is to classify documents into number of categories. In large number of labeled training document, it is difficult to create labeled training document, but easy to collect the unlabeled documents. To solve this problem unsupervised learning methods are used. Koa and Seo present the research work in this field and they propose a method that divides document into sentences, and categorized each sentence using keyword lists of each category and sentence similarity measure [12].

B. *Lexicon Approach*
Opinion words are divided in many categories. Positive opinion words are used to express some necessary things, and negative opinion words are used to describe unnecessary things. Opinion phrases and idioms are also there which together are called *opinion lexicon.* To collect the opinion word list there are three main methods. One of them is Manual method not used alone and which is very time consuming.
The basic steps of the lexicon based techniques are outlined below [22]:

1. Preprocess each text (i.e. remove noisy characters and HTML tags)
2. Initialize the total text sentiment score: s<-0.
3. Tokenize text. For each token, check if it is present in a sentiment dictionary.
   (a). If token is present in dictionary,
      I. If token is positive , then s<-s+w.
      II. If token is negative, then s<-s-w.
4. Look at total text sentiment score s,
   (a). If s> threshold, then classify the text as positive.
   (b). If s< threshold, then classify the text as negative.

I.B.1 *Dictionary Based Approach*
Manually collect a small set of opinion words and the main strategy of dictionary-based method is presented in [11, 14]. Then, this set is grown by finding their synonyms and antonyms in the WordNet [13] and thesaurus [15]. After

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

found new words these words are added to the seed list and the next process starts. This process stops when no new words are found. To remove or correct the errors manual inspection process will be done.

Disadvantage: This method cannot find the opinion words with domain and context specification orientations are the major disadvantage.

I.B.2 *Corpus Based Approach*
It is hard to prepare a huge corpus to cover all English words so it is not effective as the dictionary-based method when it used alone. But it can help to find domain and context specific opinion words using a domain corpus is the huge advantage of this method. The corpus-based approach is performed in statistical approach or semantic approach.

I.B.2.1 Statistical Approach
Statistical techniques find the co-occurrence patterns or seed opinion words. It can be done by obtaining posterior polarities in corpus, as proposed by Fahrni and Klenner [17]. By using the entire set of indexed document it is possible to solve the problem of the unavailability of some words [16].
The word has positive polarity if it occurs more frequently in positive texts, or its polarity is negative if it occurs more and more time in negative texts. If it has same occurrence, then it is neutral word. So, the polarity of word can be identified by analyzing the occurrence frequency [19].

I.B.2.2 Semantic Approach
For computing the similarity between words this method gives sentiment values directly and depends on different principles. For semantically close words this principle gives similar semantic values. By using relative count of positive and negative synonyms of this word determine the sentiment polarity of an unknown word [18].

To build a lexicon model for the description of verbs, nouns and adjectives to be used in SA and also in many other applications. The detailed subjectivity relations among the actors in a sentence expressing separate attitudes for each actor described by Maks and Vossen model [23]. Their results showed that the sometimes actor's subjectivity and speaker's subjectivity can be constantly distinguished.

To perform sentiment analysis task semantic methods can be mixed with the statistical methods. As per the Zhang and Xu [20] to find the product weakness from online reviews they used both methods. To find the frequent and infrequent denotative features they used Hownet-based similarity measure. By applying semantic methods they have grouped products feature words appropriate aspects. They took the impact of adverbs of degree and they have utilized sentence-based sentiment analysis method.

IV. CONCLUSION

The main goal of this paper is to evaluate the ensemble method for sentiment classification. This survey paper delivered an overview of recent updates in sentiment analysis and classification methods. Many of the articles cited in this paper give their contribution to the real-world application. Mine the big unstructured data has become an important research problem. Many of the organizations have putting their efforts in finding the best system for sentiment analysis. Some of the algorithms give good results but still many more limitations in these algorithms. Many of the researchers reported that SVM gives good accuracy compare to other classification techniques, but still it has some limitations. More and more future work needed because each and every organization wants to know how customers feel about their products and their services and of course about their competitors. Many different types of techniques are combined to overcome individual's limitations and benefit from each other's merit and measure the performance of classification technique.

REFERENCES

1. AggarwalCharu C, Zhai Cheng xiang," Mining Text Data", Springer New York Dordrecht Heidelberg London: Springer Science + Business Media, LLC'12; 2012.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

2. Kang Hanhoon, YooSeongJoon, Han Donglil,"Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews" Expert SystAppl ,39:6000-10,2012.
3. Ortigosa-Hernandez Jonathan, Rodriguez Juan Diego, Alzate Leandro, Lucania Manuel, InzaInaki, Lozano, Jose," A. Approaching sentiment analysis by using semi-supervised learning of multi-dimensional classifiers"Neurocomputing;92:98-115,2012.
4. Cortes C, VapnikV.,"Support-vector networks", presented at the Machine Learning; 1995.
5. Diana Maynard, Adam Funk.,"Automatic detection of political opinions in tweets."In: Proceedings of the 8th international conference on the semantic web, ESWC'11; p. 88-99,2011.
6. Vapnik V.," The nature of statistical learning theory", New York; 1995.
7. Aizerman M, Braveman E, RozonoerL.,"Theoretical foundations of the potential function method in pattern recognition learning",. Autom Rem Cont:821-37,1964.
8. Joachims T. ,"Probabilistic analysis of the rocchio algorithm with TFIDF for text categorization", In: presented at the ICML conference; 1997.
9. Chin Chen Chien, Tseng You-De,"Quality evaluation of product reviews using an information quality framework",Decis Support Syst;50:755-68,2011.
10. Li Yung-Ming, Li Tsung-Ying,"Deriving market intelligence from microblogs",Decis Support Syst,2013.
11. Hu Minging, Liu Bing," Mining and summarizing customer reviews", In: Proceedings of ACM SIGKDD international conference on Knowledge Discovery and Data Mining (KDD'04); 2004.
12. Koyoungioong, SeoJungyun,"Automatic text categorization by unsupervised learning",. In: Proceeding of COLING-00 the 18th international conference on computational linguistics;2000.
13. Miller G, Beckwith R, Fellbaum C, Gross D, Miller K,"WordNet: an on-line lexical database", Oxford Univ. Press; 1990.
14. Kim S, Hovy E," Determining the sentiment of opinions", In: Proceedings of international conference on Computational Linguistics (COLING'04);2004.
15. Mohammad S, dunne C, Dorr B," Generating high-coverage semantic orientation lexicons from overly marked words and a thesaurus" In: Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP'09); 2009.
16. Turney P.,"Thums up or thumbs down?: semantic orientation applied to unsupervised classification of reviews", In: proceedings of annual meeting of Association for Computational Linguistics (ACL'02);2002.
17. Fahrni A, KlennerM.,"Old wine or warm beer: target-specific sentiment analysis of adjectives", In: Proceedings of yhe symposium on affective language in human and machine, AISB; p. 60-3,2008.
18. Kim S, Hovy E., "Determining the sentiment of opinions", In proceedings of international conference on Computational Linguistics (COLING'04); 2004.
19. Read J, carol J.," Weakly supervised techniques for domain independent sentiment classification", In: Proceeding of the 1st international CIKM workshop on topic-sentiment analysis for mass opinion; P.45-52, 2009.
20. Zhang Wenhao, HuaXu, Wan Wei.," Weakness finder: find product weakness from Chinese reviews by using aspects based sentiment analysis", Expert SystAppl;39:10283-91,2012.
21. Y. Yang, X. Liu, "A re-examination of text categorization methods", in: Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), ACM New York, NY, USA, pp.42-49,1999.
22. M. Annett, G. Kondrak, "Acomparison of sentiment analysis techniques: Polarizing movie Blogs", In Canadian Conference on AI, pp. 25-35, 2008.
23. Maks Isa, VossenPiek. "A lexicon model for deep sentiment analysis and opinion mining applications.",Decis Support Syst;53:680-8,2012.