ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 10, October 2017

A survey on various types of Sentiment Analysis Approaches from Big data

[1] Navaneethakrishnan P, [2] Ravikumar B
[1][2] Assistant Professor ,CSE
[1][2] Velammal Engineering College -Chennai

Abstract - **In recent years, Big Data Analytics has become an essential topic for researchers.It gains more popularity due to immense data set that becomes overwhelming to users. Therefore itis essential to extract opinions from the internet and predict online customer's preferences, which could prove valuable for economic or marketing research.This inspire the researcher to develop a system that can identify and classify opinion the huge amount of text data based on the approach of Sentiment Analysis or Opinion Mining. The paper presents a survey covering the techniques and methods in sentiment analysis and challenges appear in the field. Sentiment analysis is done in data from applications like social network. There is a need for analyzing the sentiments of data and thereby defining the behaviour of the user. This involves feature extraction and thereby developing relationship trees within the scope of data.**

Keywords: Sentiment analysis, Big Data, Sentiment Classification,Feature selection.

1. INTRODUCTION

Sentiment Analysis is the standard procedure in which the entire polarity of the textual document is evaluated. It provides ameasure of how "positive" or "negative" by considering the portion of text. Sentiment analysis has applied to several applications in various fields. In Social Media perspective, sentiment analysis has significant applications in the area of monitoring, review of consumer products, campaigning and financial markets.

Development of community detection arguments, virtual communities are depends on the principle of sharing sentiments and opinions within identical groups. In this system, a logical approach is used to join the sentiment analysis and community detection, as the participants of the same communities are used to share similar sentiments.

In this context of sentiment analysis for the virtual communities, the interaction between the single sentiments of the users and the sentiment of the comprising communities is bidirectional. Determining the sentiment of the users will tends to the overall sentiment of the community, and also by learning the overall sentiment of the community tends to provide prior information for the sentiment of new members willing to join this community.

The vast amount of subjective texts present in the internet arise demand for sentiment mining. To deal this subjective data overload, in recent years, researchers have introduced various sentiment mining approaches with the goal to discover the necessary information from reviews and then provided to users. Also, the distribution imbalance of review texts on the positive and negative classes weakens the performance of classifiers. This paper presents a review covering the techniques and methods in sentiment mining and challenges appear in the field.

This paper is organized as follows. Section II presents a general overview of sentiment mining levels. Section III explains technical perspectives of sentiment classification methods. Section IV presents an overview of various application areas of sentiment analysis. Finallywe conclude the paper in Section V

2.1. Sentiment Mining Levels

Most sentiment mining approaches can be classified into three levels such as document level, sentence level and feature level. In this section, a brief review of related work on document level and sentence level sentiment mining and then discuss two representative types of methods that have been so far proposed for feature level sentiment mining.

2.1.1. Document level sentiment mining

Document level sentiment mining is to obtain an overall sentiment value for the whole document. Turney. P. D. (2002) applied point wise mutual information to calculate an average semantic orientation score of extracted phrases for determining the document's sentiment orientation. Kim.S.M. et al., (2004) calculated the document sentiment value using a self-defined formula and classified the document as positive or negative according to the value of document sentiment value. Ye.Q. et al., (2005) classified positive and negative by applying the point wise mutual information method proposed by Turney (2002) to a chinese

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 10, October 2017

document. Wilson.T. et al., (2005) developed a new sentiment mining model (OpinionFinder) that automatically identifies the appearance of opinions and sentiments, via the subjectivity analysis. The most popular sentiment learning techniques aresupport vector machine (SVM) and naive bayes (NB), and many authors have reported better accuracy by using SVM (Dang.Y. et al., (2010), O'Keefe.T. et al., (2009), Prabowo.R. et al.,(2009), Ye.Q.et al.,(2009), Abbasi.A. et al., (2008) , Pang . B et al., (2002)). Hwang.J.et al.,(2008) developed part of the sentiment word manually and by forming it into a feature vector they classified document as positive or negative with a supervised learningalgorithm. Denecke. K (2009) performed a rule based classification and a machine learning classification by using an average score and a frequency of word by class as well as a number of parts of speech of a word as a feature. Ye.Q. et al., (2009) classified a review document as positive or negative by applying a traditional topic based document classification method. Zhu.J. et al., (2010) used an artificial neural network based individual model, which showed a better classification result than either the SVM or the Hidden markov Model on Cornellmovie review data. Miao.Q. et al., (2010) calculated the ranking by analyzing the quality of a sentiment with a study on the sentiment mining and retrieval system and classified the sentiment as positive or negative using the naive bayes algorithm. Sharma.A. et al., (2012) also used artificial neural networks and sentiment lexicon for document level sentiment classification. Morae.R. et al., (2013) attempted a back propagation neural network based document level sentiment prediction for movie reviews.

2.1.2. Sentence level sentiment mining

Sentiment mining at the sentence level is another most popular approach. Besides the feasibility of automated sentence level sentiment detection, the lack of sentiment bearing features in short text units makes it very challenging. Hatzivassiloglou. V.et al., (2000) described a corpus tagged at the sentence level for subjectivity. They employed a naive bayes classifier using syntactic classes, punctuation and sentence position as features. They achieved an average accuracy of 21% more than the baseline method used in tenfold cross validation experiments by using simple features, such as pronouns, adjectives, cardinal numbers and modals. Yu.H. et al., (2003) proposed an approach to differentiate sentiments from facts at document and sentence level. Their method is based on the subjectivity scoring. The subjectivity score of a sentence is positive if most of the adjectives, adverbs, nouns and verbs in the sentence were positive. Nasukawa. T. et al., (2003) applied a methodin which the sentiment of each sentence is analyzed by identifying the sentiment expressions and subject terms. A few other works exist that perform sentence level sentiment mining (Kim.S.M. et al., (2004), Zhang.W. et al., (2007)). In Kim.S.M. et al., (2004), the sentiment words are classified individually and then the polarity of the opinion sentence is calculated by combining the individual opinion word polarity. Zhang.W. et al., (2007) proposed techniques to extract opinions contained in the web blogs. Also there exist a few product rankingmethods based on sentence level sentiment mining of product reviews for specific languages, such as Chinese(Tian.P. et al., (2009)). Zhao.J. et al., (2008) performed sentence level sentiment analysis, without focusing on the determination of representative features of the review. However, although the above works are all related to sentiment mining, researchers mainly targeted to discover the sentiment to represent a reviewer's overall opinion, but did not find which features the reviewer actually liked and disliked. For example, an overallnegative sentiment on an object does not mean that the reviewer dislikes every aspect of the object. Thus, the major focus of this research work is on the feature level sentiment mining.

2.1.3. Feature level sentiment mining

To discover a reviewer's sentiment in depth on every aspect that is mentioned in the text, some researchers have tried to mine and extract opinions at the feature level. The major task in feature level sentiment mining is to identify the individual elements which form the opinion. In recent years, the approaches for feature level sentiment mining have evolved into two principal branches: statistical based methods and model based methods.

2.1.3.1. Statistical approaches

Statistical approaches are domain independent which do not depend on labelled data and model training process.
Hu.M. et al.,(2004) proposed a popular work in which they extracted product features through association rule mining and identified opinion words that are adjacent to frequently appearing features. Later ,Popescu.A. et al., (2005) developed a opinion mining model called OPINE based on the work of Hu. M. et al.,(2004). They applied statistical analysis to find the probability that a candidate term is relevant in a product domain. Scaffidi.C. et al., (2007) also developed a sentiment mining system called Red Opal to examine consumerreviews. They identified the product features, and then scored each product on every feature. Titov.I. et al., (2008) proposed a method using latent Dirichlet allocation. In their method, they first identified

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 10, October 2017

different domains in the corpus, and then they identified the individual features of the entities in a given domain. Since all the above works are based on rules or statistical methods, their accuracy might be restricted as many of other types of entities such as non independent features, components and functions cannot be identified automatically. Thus, in recent years, many researchers have attempted to focus on supervised learning methods in order to increase the sentiment mining efficiency.

2.1.3.2. Model based approaches

Zhuang.L. et al., (2006) proposed a supervised learning method to extract the feature opinion pairs. They used annotated dataset to learn the opinion. They used a combination of dependency and part of speech paths connecting such pairs from the annotated dataset. They evaluated the algorithm on a set of movie reviews. The results yielded an improved f-score compared to the results of Hu.M. et al.,(2004). Feiguina.O. et al.,(2007) developed an approach which is independent both of search engine and the web corpus using an information extraction system. Their system learns a language model on the part of speech patterns. This was accomplished by training it on a labelled dataset. They evaluated the approach in a cross domain setting and reported the precision value of the feature extraction alone. Ferreira.L.et al.,(2008) further emphasized conditional random field's ability in addressing the cross domain applicability problem. They specifically evaluated the model's accuracy in extracting opinion targets. Kessler.J. et al., (2009) emphasized finding out feature opinion pairs within a sentence. They manually annotated features and opinions in datasets of product reviews. In their work, they provided detailed guidelines for annotation and trained a SVM classifier for extracting feature and associated opinions. Their approach reported a f-score of 0.698. Lately, Jin et al., (2009) proposed a sentimentmodel based on lexicalized hidden markov model called as OpinionMiner. This model integrates multiple important linguistic properties into an automatic learning process. Choi. Y. et al., (2009) applied conditional random fields to identify opinion holders from review data. In their error analysis, they reported that the inaccurate identification of opinions had considerable negative impact on the results. Miao.Q. et al., (2010) used conditional random fields to extract the features. They extracted features with a precision 86% in movie dataset. Nearly 15% of the product features were merged into their sentiment lexicon and in electronic product datasets, the percentage increases up to 25%. Li.S. et al., (2010) have shown that it is feasible to perform feature level review extraction based on conditional random fields, they applied it to identify single object features and opinions. Thus, in this work, the focus is not only to extend these work to address feature level sentiment mining issues, but also to integrate several improvements such as feature reduction and optimization of learning functions. In addition, in depth comparison is done among the various learning algorithms.

3. SENTIMENT CLASSIFICATION

Methods In relation to sentiment analysis, the literature survey done indicates two types of techniques including machine learning and semantic orientation. Machine learning approach applies the popularalgorithms 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.

3.1. Machine learning methods

The sentiment classification methods using machine learning approaches can be classified into supervised and unsupervised learning methods. The supervised methods make use of a large number of labelled training documents. The unsupervised methods are used when it is difficult to find these labelled training documents. Much research exists on sentiment mining of user opinion data, which mainly judges the polarities of user reviews. Machine learning approaches applicable to sentiment mining, mostly belongs to supervised classification (Biba.M. et al., (2014)). Machine learning techniques like NB, maximum entropy (ME), and SVM have achieved great success in sentiment categorization. The other most well known machine learning methods such as K-nearest neighbourhood, ID3, C5, centroid and winnow classifier are also used for sentiment mining. Among the various machine learning methods, SVM achieved great success in sentiment categorization. Naive bayes ranks next highest in performance. Naive Bayes algorithm is the widely used algorithm for document classification (Xia.R. et al.,(2011, Melville.P. et al.,(2009), Ye.Q. et al., (2009), Tan.S. et al., (2008)). The basic idea is to estimate the probabilities of categories given a test document by using the joint probabilities of words and categories. Support vector machine (SVM), a discriminative classifier is considered the best text classification method (Xia.R. et al.,(2011) , Prabowo.R. et al., (2009), Ye.Q. et al., (2009), Tan.S. et al., (2008)). Based on the structural risk minimization principle from the computational learning theory, SVM seeks a

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 10, October 2017

decision surface to separate the training data points into two classes and makes decisions based on the support vectors that are selected as the only effective elements in the training set. Multiple variants of SVM have been developed in which multi class SVM is used for sentiment classification (Xu.T. et al., (2012)). The idea behind the centroid classification algorithm is extremely simple and straightforward (Tan.S. et al.,(2008)). Initially the centroid vector for each training class is calculated, then the similarity between a testing document to all centroid is computed, finally based on these similarities, document is assigned to the class corresponding to the most similar centroid. The K-nearest neighbour is a typical example based classifier that does not build an explicit, declarative representation of the category, but relies on the category labels attached to the training documents similar to the test document. Given a test document, the system finds the K-nearest neighbours among training documents. The similarity score of each nearest neighbour document to the test document is used as the weight of the classes of the neighbour document (Tan.S. et al., (2008)). Winnow is a well known online mistaken driven method. It works by updating its weights in a sequence of trials. On each trial, it first makes a prediction for one document and then receives feedback, if a mistake is made, it updates its weight vector using the document. During the training phase, with a collection of training data, this process is repeated several times by iterating on the data (Tan.S. et al., (2008)). Besides these classifiers other classifiers like ID3 and C5 are also investigated (Prabowo.R. et al., (2009)). Besides using these above said machine learning methods individually for sentiment classification, various comparative studies have been done to find the best choice of machine learning method for sentiment classification. Tan.S. et al., (2008) presents an empirical study of sentiment categorization on Chinese documents. They investigated four feature selection methods and five learning methods (centroid classifier, K-nearest neighbour, winnow classifier, Naive Bayes and SVM) on a Chinese sentiment corpus. From the results he concludes that, SVM exhibits the best performance for sentiment classification. When applying SVM, naive Bayes and n-gram model to the destination reviews, Ye.Q. et al., (2009) found that SVM outperforms the other two classifiers. Prabowo.R. et al., (2009) described an extension by combining rule based classification, supervised learning and machine learning into a new combined method. For each sample set, they carried out 10-fold cross validation. For each fold, the associated samples were divided into training and a test set. For each test sample, a hybrid classification is carried out, i.e., if one classifier fails to classify a document, the classifier passes the document onto the next classifier, until the document is classified or no other classifier exists. Neural networks have seen a rapid growth over the years, and are being applied successfully in various application domains for the classification problems. But the state of the art techniques for neural network based text sentiment classification are found to be rare from the literature (Morae.R. et al (2013), Sharma.A. et al., (2012), Zhu. J. et al.,(2010)) . Artificial Neural Networks (ANN) are rarely been investigated in the literature of sentiment classification. In recent years we witnessed the advance in neural network methodology, like fast training algorithm for deep multilayer neural networks. Zhu. J. et al., (2010) used back propagation neural network to predict sentiment. Their model uses individual model based on ANN to determine text sentiment classification. The experimental results show that the accuracy of their model is higher than that of SVMs and hidden markov model classifiers on movie review corpus. Sharma.A. et al., (2012) also used ANN and sentiment lexicon for sentiment classification. Morae.R. et al., (2013) attempted a back propagation neural network based sentiment prediction for movie reviews. They adopted a standard evaluation context with popular supervised methods for feature selection and weighting in a traditional bag of words model. Even though Ghiassi.M. et.al. (2013) very recently used dynamic neural network for sentiment analysis and compared their model with SVM. Their focus is on twitter sentiments and they estimate the models using usual measures like recall and accuracy. They showed that dynamic ANN produces more accurate sentiment classification results than SVM. Cambria.E et.al (2013) proposed a novel cognitive model based onthe combined use of multi dimensional scaling and artificial neural networks. Though few studies exist in sentiment classification using neural networks, the literature does not contribute much work in sentiment classification using the probabilistic neural network. From the literature work done, PNN model is not applied so far in sentiment mining of product reviews to our knowledge. But many researchers have proved that PNN model is more effective than other models for data classification in various other domains (Savchenko.A.V. et al.,(2013), Adeli.H.et al.,(2009), Ciarelli.P.M.et al., (2009), Hajmeer. M. et al.,(2002). This motivates to evaluate the use of popular neural network based approach, the probabilistic neural networks (PNN).

3.2. Lexicon based approach

Since the focus of this study is on the overall sentiment (positive or negative) expressed in a review using supervised learning, this literature review is oriented

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 10, October 2017

towards supervised sentiment classification. So, a brief overview of lexicon based approach is provided. The lexicon based approach depends on finding the opinion lexicon which is used to analyze the text. There are two methods in this approach. 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 begins 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 (Balahur.A., et al., (2014), Ciarelli.P.M. et al.,(2009), Blitzer. J. et al., (2007), Kennedy. A. et al., (2005),Chaovalit. P. et al., (2005)).

3.3. Sentiment Classification on Imbalanced Datasets

Sentiment analysis has been the work of many researchers over the past years. From the literature, it is found that the machine learning approach applied mostly belongs to supervised learning for sentiment classification in balanced data distribution (Wang.G. et al., (2014), Moraes. R. et al., (2013), Tang.H. et al., (2009), Tan.S. et al., (2008)). Among classification methods used, VMs have been extensively studied and have shown remarkable success in sentiment classification applications (Xia.R. et al.,(2011), Melville.P. et al.,(2009), Ye.Q. et al., (2009) , Tan.S. et al., (2008)). Also, few studies exists on classifying customer's sentiment orientations of products using combination of classifier in balanced data distribution (Wang.G. et al., (2014), Su.Y. et al., (2013), Li.S et al., 2012, Oza N.C. et al.,(2008)).They showed that generalization ability of an ensemble method is usually much stronger than that of a single learner, which makes ensemble methods very attractive. Another characteristic of the sentiment classification literature is that many methods have been tested only on balanced datasets and there has been little discussion on the effects of learning subjective aspects from imbalanced data, although it is typical of the product domain to have substantially more positive than negative reviews (Li.S. et al., (2012), Burns .N. et al., (2011), Wang.S. et al., (2011)). Burns.N. et al., (2011) address sentiment classification on imbalanced data, however the experiments do not involve SVM. Li.S. et al. (2011a, 2011b) adopted a random under sampling method, which is a popular approach to deal with imbalanced data. The major drawback of random under sampling is that it can discard potentially useful data that could be important forthe learning process. To overcome this, Wang.S. et al., (2011) proposed combining multiple classifiers, which are trained from multiple instances of under sampled data. However, the real time sentiment analysis is a challenging machine learning task, due to the imbalanced nature of positive and negative sentiments. This motivates to deal with imbalanced datasets for sentiment classification. Sentiment analysis becomes complex. when learning from imbalanced data sets, very few minority class instances cannot present sufficient information and result in performance degrading significantly. Modifying the data distribution or the classification algorithm is the traditional approaches to dealing with the class imbalance problem in other application areas of research (Chawla.N. et al., (2002), Sun.A. et al., (2009)). In this work, a new method is proposed using the combination of both approaches. A modification is proposed in bagging ensemble algorithm and also in sampling method used for data distribution, so as to solve a class imbalance problem to improve the classification performance.

3.4. Feature Selection

Sentiment mining by machine learning methods consists of two steps: (1) features extraction from training data and converting them to feature vectors and (2) training of the classifier on the feature vectors and testing the instances. Transforming a text message into a feature vector is an important part of machine learning methods for sentiment classification. The features obtained from text messages should include only relevant information and also to be independent of each other. It can make classifiers more efficient by reducing the amount of data to be analyzed as well as identifying relevant features to be considered in the learning process. Various feature selection methods are applied in the sentiment classification literature such as document frequency (Dang.Y. et al., (2010), Tan.S. et al., (2008), Pang . B. et al., (2002)), mutual information (Tan.S. et al., (2008), Li.T. et al., (2008), Turney.P. D. et al., (2002)), information gain (Abulaish . M. et al., (2009), Tan.S. et al., (2008),Abbasi. A. et al.,(2008b)) and chi square (O'Keefe.T. et al., (2009), Ferreira. L. et al.,(2008), Tan.S. et al., (2008), Gamon.M. (2004)). None of them has been widely accepted as the best feature selection method for sentiment classification. However, information gain has often been competitive. It ranks terms by considering their presence and absence in each class. Another feature representation method dominating the sentiment classification literature is known as the bag of words framework. In this framework, the text message is considered as a bag of words and represented by a vector containing all the words appearing in the corpus. Besides bag of words features, many other types of features are proposed, such as part of speech, syntax, negation, and topic oriented features (Yuan.M. et al., (2013)). However, these

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 10, October 2017

features rely heavily on linguistic resources (Xia.R. et al., (2013), Wu. C.H., et al., (2006), Popescu.A. et al., (2005), Wilson.T. et al.,(2005a), Pang . B et al., (2002), Turney. P. D. (2002)). Pang.B et al. (2002) used syntactic features (unigrams, bigrams, unigrams + POS, adjectives, and unigrams + position). Yu.H et al., (2003) used words, bigrams, and trigrams, as well as the parts of speech as features in each sentence. Gamon.M (2004) presented a feature reduction technique based on log likelihood ratio to select the important attributes from a large initial feature vectors. Wilson.T. et al., (2005a) used a hierarchy to define different types of lexical features and their relationship to one another, in terms of both representational coverage and performance. They proved that the reduced feature set can improve the performance on three sentiment classification tasks, especially when combined with traditional feature selection approaches. Wilson.T. et al., (2006) used collocation technique where certain parts of fixed-word n-grams were replaced with general word tags, thereby also creating n-gram phrase patterns. Wang.S.G. et al., (2007) presented a hybrid method for feature selection based on the category distinguishing capability of feature words and information gain.Tan.S. et al., (2008) presented an empirical study of sentiment classification on Chinese documents. They compared four feature selection methods such as information gain, mutual information, chi square and document frequency. The experimental results indicate that information gain performs the best for selecting the sentimental terms for Chinese product domain. Sharma.A. et al., (2012) and Singh.V.K. et al., (2013) introduced manual or semiautomatic approaches for generating sentiment lexicons that uses an initial set of automatically generated terms which are manually filtered and coded with polarity and intensity information. The user defined tagsare incorporated to indicate whether certain phrases convey positive or negative sentiment. Yu. L.C. et al., (2013) used semi automatic lexicon generation tools to construct the sets of strong subjectivity, weak subjectivity, and objective nouns. Their approach outperformed the use of other features (e.g. bag of words) for objective classification. For the very noisy domain of customer feedback data, feature reduction methods are also applied to obtain a reduction of the original feature set by removing some features that are considered irrelevant for sentiment classification to yield an improved classification accuracy of learning algorithms. Wang, S. et al., (2011) employed a hybrid method for feature selection and fisher's discriminant ratio for feature reduction. But the literature does not contribute much work using popular feature reduction method, principal component analysis in sentiment classification (Cambria. E., et al., (2013)). Thus feature selection and reduction are crucial problems that have been tackled by many researchers in different ways. However, no consistent conclusions have been found from these studies that one technique is superior to other (Swaminathan et al., (2010), O'Keefe.T. et al., (2009), Abbasi. A. et al.,(2008b),Salton.G. et al., (1997)). Since discussions of effective feature selection and reduction are beyond the scope of this research, in this study the most widely used unigram, bigram and trigram features are adopted and popular principal component analysis (PCA) is applied as feature reduction method.

4. APPLICATIONS

Sentiment mining has various applications in different fields. It can be used in recommendation systems, search engines, web advertisement filtering, email filtering, questioning-answering systems, etc. Sentiment mining application in daily life is most interesting as this can be used to improve man machine interactions, business intelligence, government intelligence, citation analysis etc. Some authors have specifically worked on applications for customer reviews (Balahur. A et al (2012) , Ganesan.K et al., (2010), Jin et al., (2009), Thet.T et al., (2007), Chen.L.S et al., (2006), Liu.B et al., (2005)), while others have applied sentiment mining to the mining of newspapers and websites in to extract public opinion (Liang.J.et al.,(2014),Maragoudakis.M. et al.,(2011),Stepinski.A.et al., (2007)).Ghani.R. et al., (2004) has applied the concept of sentiment mining to online auctions to predict the end price of items, while other papers have reported work on public opinion mining for government decision making (Vonitsanou.I.K. et al., (2012)). Furuse.O et al., (2007) developed an open domain query based search engine for extracting statements of sentiment. Although sentiment mining can be applied to the social and business sectors, researchers are also making an effort to effectively employ it in other important areas, e.g., health, education, travel, restaurant etc. Goeuriot. L. et al.,(2011) proposed social media sites where people post information about their diseases and treatments for the purpose of mining disease and treatment information. In an interesting application of sentiment mining, Swaminathan et al., (2010) extract relationships between bio entities, such as food and diseases. Xia.R et al., (2009) and Sohn.S. et al., (2012) applied sentiment mining techniques to classify opinions in medical domain. Furthermore, sentiment mining is being applied in several commercial areas such as tourism, automobile purchasing, electronic product reviews, movie

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 10, October 2017

reviews, and game reviews as well as in various political arenas such as public administration, strategic planning, marketing etc. (Yuan.M. et al., (2013), Abulaish.M et al.,(2009), Blitzer.J et al., (2007), Feldman.R et al., (2007), Kessler.J. et al., (2010), Zhuang.L. et al., (2006), Zhang.Z. et al., (2011)). The above mentioned works represent only a small sample of sentiment mining applications. Various potential applications of sentiment mining seen from this survey of the existing works indicates the importance of sentiment mining in practical life (Himmat.M. et al., (2014), Tang.S et al., (2009), Pang .B et al.,(2008)).

5. CONCLUSION

On the strength of the exhaustive review of work done by previous researchers, it was found that a good amount of work was done on machine learning methods. However, it was evident that most of the work had been done on balanced datasets. From the literature, it was also observed that, the research focus was not much on using imbalanced datasets and on analyzing the hybrid combination of classifier with PCA. Thus motivates to analyze the effect of classifiers, both balanced and imbalanced datasets. From the literature done, the following are the main objectives of the present investigation to expand on existing solutions used for automatic sentiment mining.

S.No	Author Name	Technique used	Feature Selection	Data Source
1.	Tan. S et al (2008)	Centroid, KNN, SVM	mi, ig, chi, df	Chnsenticorp
2.	Melville. P et al (2009)	Bayesian Classification	N-Grams	Blog post
3.	Jin et al (2009)	Id3-negation Phrase	Symmetric	Mobile Review
4.	Miao.Q et al (2010)	Dictionary based approach	Apriori	Amazon
5.	Zhu. J et al (2010)	Backpropagation	Unigram	Movie Review
6.	Zhang et al (2011)	Naïve Bayes , SVM	Character based bigrams	Cantonese review
7.	Sheng et al (2011)	BPN	SO approach	Movie Review
8	Yanghui Rao et al (2014)	Naïve bayes	N-grams	Movie Review
9	Alvaro et al (2014)	SVM	Symmetric	Mobile Reveiw

REFERENCE

[1]. Zhi-Hong Deng , Kun-Hu Luo and Hong-Liang Yu(2014), “A study of supervised term weighting scheme for sentiment analysis”, Elsevier, Expert Systems with Applications ,Vol.41(7),PP. 3506–3513.

[2].Alvaro Ortigosa, José M. Martín and Rosa M. Carro(2014),” Sentiment analysis in Facebook and its application to e-learning”, Elsevier,Computers in Human Behavior ,Vol.31, PP.527–541.

[3]. Li.T.,Zhu.S., &Ogihara.M. (2008), “Text ategorization via generalized discriminant analysis”, Information Processing & Management, 44(5), 1684-1697. 167

[4]. Liang. J., Liu. P., Tan. J., & Bai. S. (2014), “Sentiment Classification Based on AS-LDA Model”, Procedia Computer Science, 31, 511-516.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 10, October 2017

[5]. Liu.B.,&Zhang.L. (2012), "A survey of opinion mining and sentiment analysis", In Mining Text Data, Springer US, 415- 463.

[6]. Miao. Q., Li. Q., & Zeng. D. (2010), "Mining fine grained opinions by using probabilistic models and domain knowledge", Proceedings of the 2010 IEEE/ WIC/ACM international conference on web intelligence and intelligent agent technology – WI-IAT'10 ,Washington, DC, USA: IEEE Computer Society 01, 358–365,.

[7].Min-Chul Yang and Hae-Chang Rim(2014), "Identifying interesting Twitter contents using topical analysis", Expert Systems with Applications ,Elsevier,Vol.41(9),PP. 4330–4336.

[8]. Moraes. R., Valiati. J. F. &GaviãONeto. W. P. (2013), "Document-level sentiment classification: An empirical comparison between SVM and ANN", Expert Systems with Applications, 40(2), 621-633.

[9]. Nasukawa. T., & Yi. J. (2003), "Sentiment analysis: Capturing favorability using natural language processing", In Proceedings of the 2nd international conference on Knowledge capture, ACM, 70-77.

[10]. Xia.R.,Zong.C., Hu.X., &Cambria.E. (2013), "Feature ensemble plus sample selection: domain adaptation for sentiment classification", Intelligent Systems, IEEE, 28(3), 10-18.