A MACHINE LEARNING BASED SENTIMENT ANALYSIS BY SELECTING FEATURES FOR PREDICTING CUSTOMER REVIEWS

Nagamanjula R,
Research Scholar,
Mother Teresa Women's University,
Kodaikanal.
Email: nagabtl@gmail.com

Dr. A. Pethalakshmi,
Associate Professor and Head,
M.V.Muthiah Government Arts
College for Women, Dindigul.
Email: pethalakshmi@yahoo.com

Abstract- Nowadays people can express their opinions and views publicly which can be favour and/or against any service, issue, product, event, or policy. With the rapid advancement of internet, people can share their feedback on the web in huge numbers. This large number of reviews for individuals can be crucial to improve their services and products, called Opinion Mining. It is also known as sentiment analysis which ultimate goal is to differentiate the emotions expressed within the reviews. In this paper, we analysed these reviews and classifying them into positive or negative opinions. This paper presents a novel classification method called Support Vector Machine (SVM) in order to improve the accuracy by forming two classes i.e. positive and negative. Initially, words are collected from social site Amazoon.in. (Reviews about electronic products, especially for mobile brands) and pre-processed using wordnet tool. To classify the review/comment, we selected some of the features which are evaluated from user review comments. The Information gain with Fast Correlation based Filter (FCBF) considered for feature selection and then the SVM classifier is find the classes. An intensive experimental study shows the efficiency of these enhancements and shows better performance in terms of precision, recall and f-measure.

Keywords – Opinion Mining, Sentiment Analysis, Reviews/Comments

I. INTRODUCTION

Opinion mining is defined as the process of information retrieval and widely used in web data analysis. It helps for any organization to make the decisions. The exponential growth of the user generated content has opened new horizons for researchers in the field of sentiment analysis. Sentiment analysis defines a procedure of mining, analysing, classifying and describing the sentiments or feelings in the form of word-based data using Natural Language Processing (NLP), Machine Learning (ML) or Statistics [1]. The two terms opinion mining or sentiment analysis are more substitutable. Opinion mining mine the textual data and computes public's attitude around an object whereas sentiment analysis classifies the sentiment articulated in a script then evaluates it. Sentiment analysis can be categorized in to three main categories: Aspect-level SA, Document level SA, and Sentence-level SA. 1). Aspect-level SA: This type of SA categorizes the sentiments through feature to the appropriate feature of objects [2-3]. The main objective is to classify the objects and their features. For e.g. "The camera of this phone is not good, but the voice clarity is excellent. 2). Document-level SA: This type of SA is to classify an attitude text as articulating a positive or negative attitude or sentiment. It considers the complete text as a basic data unit and 3). Sentence-level SA: This type of SA is to categorize sentiment articulated in individual sentence. Firstly, it classify the sentence is neither subjective nor objective. If the sentence becomes subjective, sentence-level SA will make the decision whether the sentence decides as a positive or negative feelings [4-6].

In online social network (OSN) sites such as Amazon, Flipkart, etc. are stored user opinions and reviews for any product or seller. This aids for new buyer to choose their products or sellers in their sites. In general, social media sites have become essential tools for people to express their opinions about their daily lives and exchange with others towards several topics. SA can be widely classified into two categories such as Lexicon based approach (Linguistic) and Machine learning based approach (Statistical). The lexicon based approach requires the construction of an opinionated lexicon of labelled words to find the semantic orientation (valence of polarity) of large text. There are five types of lexicon based approaches were presented such as manual method, dictionary based approach, corpora based method, hybrid method and concept based method whereas the statistical approach have focuses on using machine learning algorithms namely support vector machines (SVMs), naïve bayesian classifiers, neural networks and maximum entropy to identify and classify the class labels for texts [7-9].

From the performance results and comparison made by the lexicon and machine learning based approaches, the machine learning based approach gives better capability than the lexicon based approach. However, machine learning based classification techniques have some limitations due to the selection of relevant features (best features) that truly captures the sentimental and semantic features or characteristics of texts to attain the best results in classification phase. In machine learning based classification methods, feature selection is still a crucial task. Most of previous methods follow using bag-of-words (BOG) feature representation. The texts can be represented for instance as binary (absence/presence of lexicon words), TF-deltaIDF (opinionated positive and negative words), frequency (number of occurrences), weighted by their TF-IDF score (using lexicon words). However, it seems difficult to define a word with a feature vector that covers all the words that present in the language. Particularly, in data gathered from social media site, the vocabulary is very dynamic and rapidly varied. Therefore the handling of new words is very difficult [10]. Due to this reason we presented this research paper.

This research proposes a text processing through SVM method with Classification. Feature selection was used to select the relevant feature over the dataset in order to get a better performance of SVM as a classifier. In this study, the results of positive and negative sentiments are based on the extracted features from the method of information gain with Fast Correlation based Filter (FCBF).

The paper is organized as follows: We review in the next section the related work for sentiment analysis using machine learning-based methods. We then explain on the principle of our proposed method for sentiment classification in Section 3. Next in section 4, the experimental results. Finally, we conclude the paper with future works.

II. RELATED WORK

Various research related to sentiment analysis is discussed using certain methods to optimize performance levels including enhancing accuracy and reducing the error rate in classification.

In this section, we discuss an overview of different methods proposed to perform sentiment analysis task using machine learning algorithms such as supervised methods and unsupervised methods, in which sentiment analysis is stated as a classification task. It invokes opinions classifications in text into two different categories, like positive or negative. The machine learning algorithms are Support Vector Machines [11], Maximum Entropy [12], Naïve Bayes Classifier [13], Softmax [14], etc. This type of machine learning methods need a set of well-classified sentences (manually labelled data-training corpus). In general, supervised methods aim to discover a model using labelled examples which should be capable to generalize the classification learned on a wider dataset. Then it comes to learning a machine how to assign a class to a new unlabelled data among the relevant classes in the prediction model. To perform machine learning, it requires converting the text into numerical representations that may lead to accurate classification. In machine learning methods, different types of features are used in order to construct feature vector representations of text. Two different types of features are extracted: namely the bag of words vector representation (considers the order of their appearance) and the sequential representation (preserves the order of words consisting in a sentence). With the rapid development of multimedia technology, it has resulted in the rapid growth of population. Sentiment classification based on body sensor networks was introduced in [15] using distance and smaller interclass distance resulting in the improvement of sentiment classification. However the classification performance was not considered and implemented. In [16] AdaBoost Neural network was applied in automated sentiment image classification. However this method consumes more execution time and human resources.

Direct and indirect discrimination was applied in [17] to improve the true positive rate using legitimate classification rules. However this method tried to explore the relationship between discrimination prevention and privacy preservation. In [18] business intelligence (BI) was improved by applying suggestive reviews with the objective of minimizing the preprocessing time. Due to the insufficient sample size (training corpus), this model was impossible to classify the reviews within suggestive family. Intrinsic and extrinsic domain relevance was introduced in [19] with the aim of improving the extraction rate for feature candidates. It was less successful while dealing with the specific feature terms commented over explicitly in reviews.

In this study, we proposed the use of a Machine Learning Support Vector Sentiment Classification method to classify the opinion words from user review comments (amazon.in).

III. RESEARCH METHODOLOGY

In this section, we explore about the proposed system with the support of sentiments for sentiment classification. In common, the steps performed in this study comprised of determining issues about, collecting data retrieved from social networks, Text pre-processing, feature selection and classification. Figure 1 shows the architecture diagram of machine learning based sentiment analysis approach with the features selected using Information Gain and Fast Correlation based Filter (FCBF). The parts of text processing are included in table.1.

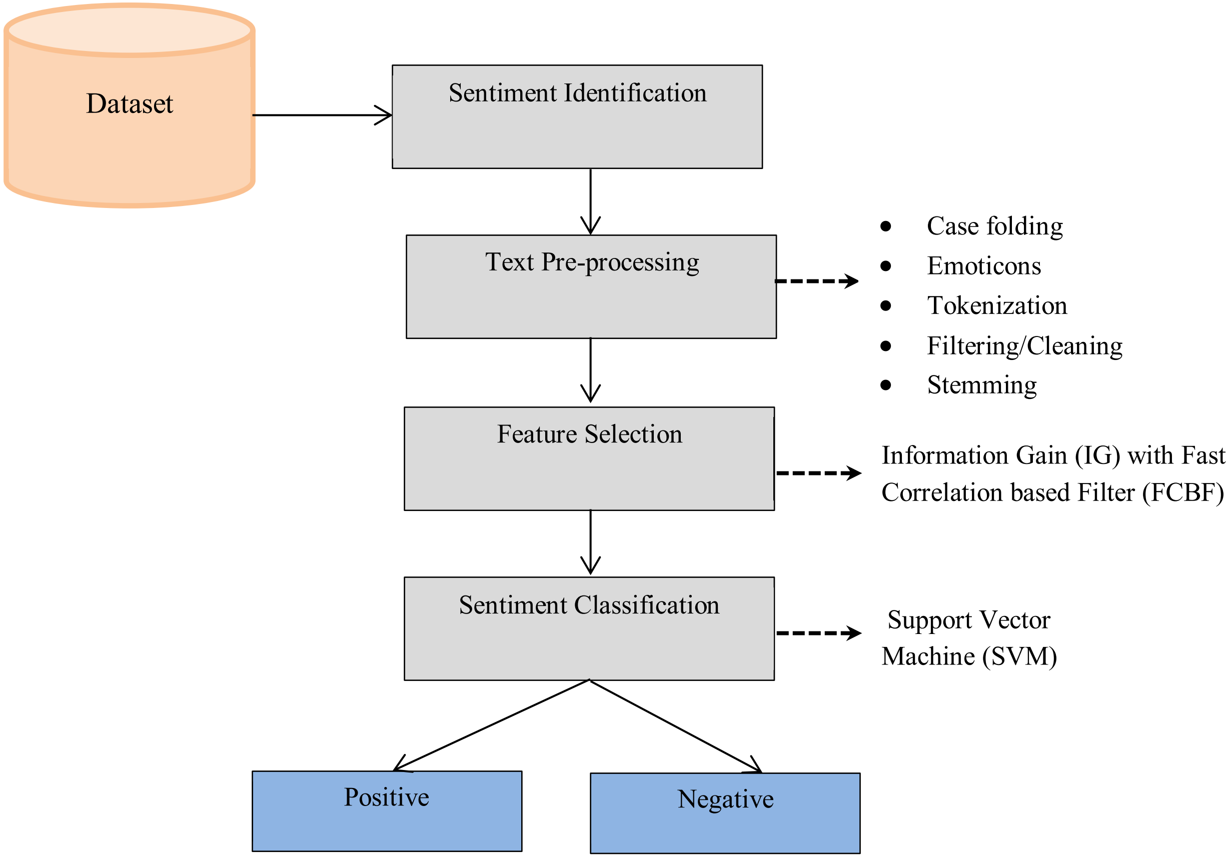

FIGURE.1. ARCHITECTURE DIAGRAM FOR PROPOSED SYSTEM USING SUPPORT VECTOR MACHINE

Table.1. Processing Requirements and Analysis

Phases	Description
Text collection	Collect reviews from social networks
Case folding	Convert the word on tweets into lowercase and by cleaning the hashtag or symbol
Emoticons	Change emotions into words to be interpreted
Tokenization	Break a sentence in to a snippets
Filtering/Cleaning	Remove or avoid any irrelevant words
Stemming	Reduce every word to get the word base
Feature selection	Select the relevant features from dataset
Classifier	Classify comments in positive and negative based on type

3.1. Data Collection Phase

In data collection stage, dataset is used to manage as a training data and data testing. After completing data collection, the next phase is to do text processing on the training dataset. Text processing is done by five stages: case folding, emoticons, tokenization, data cleaning and stemming. This process is done to regulate the structure of words received from social networks in order to facilitate the classification using sentiments.

3.2 Feature Selection Phase

This phase is used the concept of "Information gain" which is one of the famous technique used for feature selection criteria for classification of text. In general, classification of text can be used in various applications such as information extraction and information retrieval. The major advantage of information is based on the theory of information theory by computing how much class label information is retrieved when observing the value of some features. In information theory, the Entropy value is comprised in different distributions includes the distribution of class P(c). The information Gain of some features f measures the value of Entropy P(c) that modifies after observed f as the following equation:

$$IG(w) = -\sum_{c\,\epsilon C} P(c) \log P(c) + \sum_{w\epsilon(0,1)} P(w) \sum_{c\epsilon C} P(c|w)\ log\ \ P(c/w) \qquad (1)$$

In the information gain, the relevant features can be shown when the feature has a high value of information gain. In this paper, we proposed Fast Correlation based Filter (FCBF) to find the relevant features and redundant features. This approach is effective to handle both feature redundancy and relevancy during feature selection. The proposed filter worked well to select the set of feature and identify the high correlation to the class with symmetrical uncertainty (SU)

$$SU(\rho) \geq \rho = 2\frac{IG\ (f,C)}{H(f)+H(C)} \qquad (2)$$

3.3. Sentiment Analysis using SVM

SVM (Support Vector Machine) is a machine learning technique, especially designed for classification. It is highly effective at conventional text categorization schemes. It looks for a decision hyperplane with largest margin and divided the datasets into two types. SVM is capable of working on high-dimensional datasets using "Trick Kernel". In SA, support vector machine describes the sentiment polarities of articles based on feature vector that is composed of sentiment features. The feature vector can be stated as follows F = $\{f_1, f_2, \dots, f_m\}$ a set of features extracted from the training dataset. It is common to use sentiment phrases within the training dataset as a sentiment feature f_i in the feature vector. Though there are other desired ways to represent sentiment features. Most researchers tend to use sentiment phrases as features since they are basic symbols of sentiment information. For each sentiment feature, we need to measure it with a certain weight value. There are different methods for weighting feature phrases include document frequency (DF), Term Frequency-Inverse Document Frequency (TF-IDF), and Feature Presence (FP). The feature phrase with big frequency difference in two kinds of balanced corpus makes much contribution to sentiment classification. For this purpose, it has poor classification ability and must be abandoned. Our proposed method called as Relative Document Frequency (RDF). It is expressed as follows:

$$RDF(t) = |DF_1(t) - DF_2(t)| \qquad (3)$$

Where$DF_1(t) - DF_2(t)$ represents the frequency of feature t in document 1 and document 2. n_i(d) is amount of feature f_1 occurring in training corpus and the support vector for sentiment analysis in SVM is written as

$$\vec{d} = \{n_1(d)_, n_2(d)_{\dots\dots,,} n_m(d)_,\} \qquad (4)$$

If the support vector ($\vec{d}$) in SVM, general SVM tokenize each article in training corpus with a tokenizer and extract sentiment phases. Next, train the support vectors $\vec{d}$ based on the training corpus to increase the accuracy in sentiment classification for training corpus and then test the support vector on test corpus. We call this method as general SVM. Now we propose our SVM which enriched in terms of multiple objectives. In this method, we added a user dictionary to tokenizer and construct a sentiment lexicon, next we construct another sentiment vector which is contains of features in sentiment lexicon. It can be defined as $S = \{S_1, S_2, \dots, \dots, S_m\}$. S_i is a sentiment phrase in the lexicon $w_1(c)$ is the weight of the feature S_1 in the lexicon and the n we get another support vector $\vec{c} = w_1(d),.\ w_2(d), \dots,.\ w_m$. our proposed method support vector can be expressed as:

$$\vec{V} = \vec{d} + \varepsilon\vec{c} \qquad (5)$$

Where ε is amplification factor which can increase the proportion of $\vec{c}$ in$\vec{V}$. We train the vector $\vec{V}$ with a training corpus and the use it in the test corpus. $\vec{d} = \{n_1(d)_, n_2(d)_{\dots\dots,,} n_m(d)_,\}$ and $\vec{c} = \{w_1(d),.\ w_2(d), \dots,.\ w_m\}$

IV. RESULTS AND DISCUSSION

This section describes about the experimental results and the performance for proposed vs. existing systems are identified and plotted as graphs for the implemented methodologies

(a). Social Media Dataset

A training corpus considered in this paper is of mobile category of ten different mobile brands and in each mobile category, there are certain reviews. These reviews collected form the online shopping site Amzon.in and it is stored and managed in structured form [20]. It comprised of reviewer name and corresponding reviews. The data fields in this dataset are follows ten different mobile brands (Micromax, iPhone, Samsung, etc.) Around 300 reviews were stored for each mobile category. In this work, a Java application on NetBeans platform can be used. Sentiment analysis data obtained from Amazon.in can be displayed in Table 2.

Table.2. Number of Reviews on Different Mobile Brands

Mobile Brand Names	Category	Number of Reviews	Positive	Negative
Apple iPhone 6	Electronic	300	154	146
Sony Xperia M2	Electronic	250	151	99
Samsung Galaxy S6	Electronic	240	178	62

According to the data in table 2, this paper retrieved text from amazon for Apple iPhone 6 is

Table.3. Amazon Datasets Results

Type of product	***Target Sentiment***	***Precision (p)***	***Recall (R)***	***F-measure***
Apple iPhone 6	Positive	0.88	0.77	0.82
	Negative	0.81	0.77	0.79
Sony Xperia M2	Positive	0.88	0.82	0.85
	Negative	0.89	0.82	0.85
Samsung Galaxy S6	Positive	0.82	0.79	0.80
	Negative	0.85	0.79	0.82
Proposed system results (average)		0.86	0.79	0.82

300 ("Positive → 154, Negative → 146"), Sony Xperia M2 ("Positive → 151, Negative → 99"), and Samsung Galaxy S6 ("Positive → 178, Negative → 62"). Various product comments have been done by text processing and feature selection and then these comments classified into Positive and Negative. The kind of features which are used for this online products like Uniterm words (Price, Battery, Processor, Display, etc.) and others about 250+ selected features. After sentiments classified, evaluation of each sentimental category for each product, particularly identify the use of proposed SVM. Experiment is conducted on the features for classification of sentiments using user review comments.

The performance metrics to evaluate the performance of proposed system are precision, recall and f-measure. Among these parameters, precision (p) can be defined as the percentage of predicted data as positive is correct or how much reviews relevant. Recall (r) is defined as the percentage of the predicted as positive or how much the selected reviews. Finally the f-measure value can be defined by the combination of precision-recall curve that finds the results of evaluation. The table of performance metrics such as precision, recall and f-measure from amazon social networking can be shown in following Table 3.

According to data in Table 3, the highest precision, recall and f-measure values in Sony Xperia M2 is 88%, 82%, and 85% and positive sentiment respectively. Besides the average of all precision, recall and f-measure from amazon were 86%, 79%, and 82% respectively. For the precision, recall, and f-measure calculation results from product data are shown in figure 2, 3, and 4.

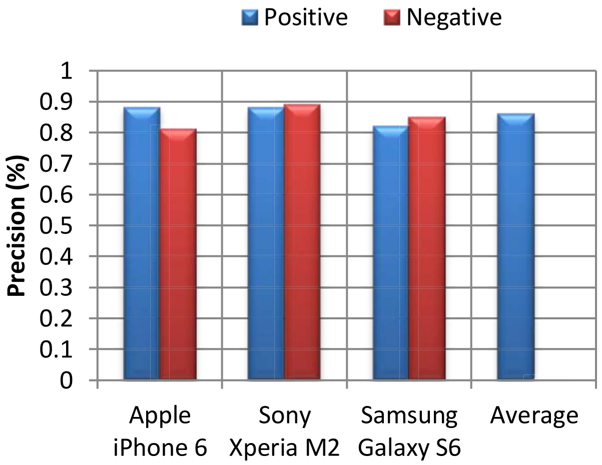

FIGURE.2. RESULTS FOR PRECISION

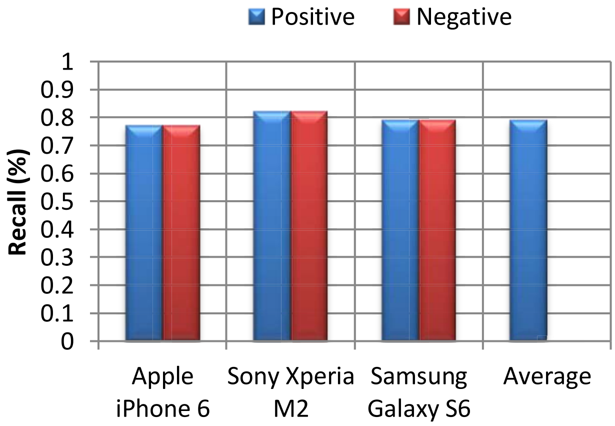

FIGURE.3. RESULTS FOR RECALL

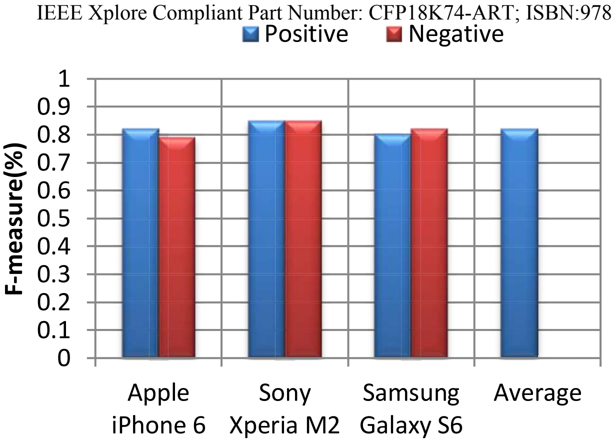

FIGURE.4. RESULTS FOR F-MEASURE

V. CONCLUSION AND FUTURE ENHANCEMENT

In our work, we used the feature selection method which identifies the opinions of product from customer reviews. This proposed SA architecture encloses four main phases, i.e. data collection, text pre-processing, feature selection and classification. Feature selection approach helps us to select the most relevant features to be considered for classification. In this work we will considered about feature selection, future work will be consider implicit and explicit aspects extraction for accurate classification and also try to improve classification performance by using hybrid machine learning algorithms.

References

1. Hanen Ameur, Salma Jamoussi, Abdelmajid Ben Hamadou, “Sentiment Lexicon Enrichment using Emotional Vector Representation”, IEEE/ACS 14th International Conference on Computer Systems and Applications, 2017.

2. Shubha, Suresh, “An Efficient Machine Learning BayesSentiment Classification Method based Review Comments”, IEEE, 2017

3. Ekki Rinaldi, Aina Musdholifah, “FVEC-SVM for Opinion Mining on Indonesian Comments of YouTube Video”, International Conference on Data and Software Engineering (ICoDSE), 2017.

4. Melva Hermayanty Saragih, Abba Suganda Girsang, “Sentiment Analysis of Customer Engagement on Social Media in Transport Online”, International Conference on Sustainable Information Engineering and Technology (IET), 2017.

5. Penubaka, Balaji, D. Haritha, O. Nagaraju, “An Overview on Opinion Mining Techniques and Sentiment analysis”, International Journal of Pure and Applied Mathematics volume118, 2018.

6. N. Sathya Priya, C. Akila, “A Survey on Opinion Mining Techniques and Online Reviews,” International Journal of Scientific Development and Research (IJSDR), Vol. 1, PP. 70-74, 2016.

7. Chetashri Bhadane, Hardi Dalal, Heenal Doshi, “Sentiment Analysis: Measuring Opinions”, Science Direct, PP. 808-814, 2015.

8. Gagandeep Singh, Kamaljeet Kaur Mangat, “Performance Analysis of Supervised Learning Methodologies for Sentiment Analysis of Tweets”, International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 5, Pp. 500-509, 2015.

9. Preety, Sunny Dahiya, “Sentiment Analysis using SVM and Naïve Bayes Algorithm”, International Journal of Computer Science and Mobile Computing”, Vol. 4, Pp. 212-219, 2015.

10. Ali Hasan, Sana Moin, Ahmad Karim, and Shahabodd in Shamshirband, “Machine Learning-based Sentiment Analysis for Twitter Accounts”, Mathematical and Computational Applications, vol. 11, Issue 23, 2018.

11. Ahmad, Munir, & Shabib Aftab. “Analyzing the Performance of SVM for Polarity Detection with Different Datasets”, International Journal Modern Education and Computer Science. DOI: 10.5815/ijmecs.2017.

12. Amir Hamzah, Naniek Widyatuti, “Opinion Classification using Maximum Entropy and K-Means Clustering”, International Conference on Information and Communication Technology and Systems, 2016.

13. Jadon, E., Sharma R. “Data Mining: Document Classification using Naive Bayes Classifier”. International Journal of Computer Applications Volume 167 - No. 6, June 2017.

14. Bhavish Khanna N, Sharon Moses, Nirmala M, “Softmax based User Attitude Detection Algorithm for Sentiment Analysis”, 6th International Conference on Smart Computing and Communications, ICSCC 2017, 2017.

15. WeiWang and Xiaodan Huang,”Divisibility and Compactness Analysis of Physiological Signals for Sentiment Classification in Body Sensor Network”, Hindawi Publishing Corporation International Journal of Distributed Sensor Networks, Article ID 937163, 11pages, Volume 2013.

16. Jianfang Cao, Junjie Chen, and Haifang Li, “An AdaBoost Back propagation Neural Network for Automated Image Sentiment Classification”, Hindawi Publishing Corporation, Scientific World, Volume 2014, August 2014.

17. Sara Hajian and Josep Domingo-Ferrer, “A Methodology for Direct and Indirect Discrimination Prevention in Data Mining”, IEEE Transactions on Knowledge and Data Engineering, Volume 25, Issue 7, Pages 1445-14593, July 2013.

18. Atika Qazi, Ram Gopal Raj, Muhammad Tahir, Erik Cambria, and Karim Bux Shah Syed, ”Enhancing Business Intelligence by Means of Suggestive Reviews”, Hindawi Publishing Corporation, The Scientific World Journal, Article ID 879323, 11 pages, Volume 2014.

19. Zhen Hai, Kuiyu Chang, Jung-Jae Kim, and Christopher C. Yang, "Identifying Features in Opinion Mining via Intrinsic and Extrinsic Domain Relevance", IEEE Transactions on Knowledge and Data Engineering, Volume 26, Issue 3, Pages 623-634, March 2014.

20. T. Sangeetha, N. Balaganesh, Muneeswari, "Aspects based Opinion Mining from Online Reviews for Product Recommenation", International Conference on Computational Intelligence in Data Sciences (ICCIDS), 2017.