An Unsupervised Fuzzy Clustering Method for Twitter Sentiment Analysis

Hima Suresh
Research Scholar, School of Computer Sciences
Mahatma Gandhi University
Kottayam, India

Dr. Gladston Raj.S
Head, Department of Computer Science
Govt. College, Nedumangadu
Trivandrum, India

***Abstract*— Cluster based techniques on sentiment analysis is a novel approach for analyzing sentiments expressed in social media sites. It is a main task of exploratory data mining, and a common technique used in machine learning. In contrast to supervised learning technique, the cluster based techniques produce essentially accurate experimental results without manual processing, linguistic knowledge or training time. This paper presents a novel fuzzy clustering model to analyze twitter feeds regarding the sentiments of a particular brand using the real dataset collected over a period of one year. Then a comparative analysis is made with the existing partitioning clustering techniques namely K Means and Expectation Maximization algorithms based on metrics namely accuracy, precision, recall and execution time. According to the experimental analysis, the proposed approach is tested to be practicable in performing high quality twitter sentiment analysis results.**

Keywords*— *Sentiment Analysis (SA); partitioning clustering techniques; Expectation Maximization (EM); Simple K- Means.

I. INTRODUCTION

Twitter is one of the most popular micro-blogging web sites in which users could send and receive a short 140 character messages called tweets. It provides means for empirical analyzing properties of interactions with people. The Information extracted from twitter could be the opinions relating to different topics such as politics, brand impact, election etc. With the advent of machine learning techniques; decision makers could ensure efficient solutions for a plethora of problems.

Many researchers have attempted to find out a technique to automatically analyze the sentiment orientation of documents especially from reviews, blogs etc. This could be categorized into two different machine learning techniques such as (i) Supervised machine learning technique and (ii) Unsupervised machine learning technique. In spite of the fact that Supervised machine learning technique enjoys a relatively high efficiency compared to Unsupervised machine learning technique, its processing requires manual participation. Unsupervised machine learning on the other hand, do not demand manual involvement but its accuracy could be limited.

This paper focuses on the twitter sentiment analysis of a brand using Partition based clustering techniques.

Partition based clustering technique is an Unsupervised machine learning technique. Two such techniques (K Means and EM) are analyzed with the proposed method regarding the brand information collected from tweets and observed that the proposed fuzzy clustering method provides better results based on the aspect of accuracy and execution time over the other two partition based techniques.

Main contributions in our work include collection of real data sets of 300 samples of tweets from Twitter API over a period of one year from 2015 to 2016 regarding the particular brand called Samsung Galaxy S6. Then a modified fuzzy clustering method has been proposed and attempted a comparative analysis with the existing partition based methods namely K means and Expectation Maximization methods.

The remainder of this paper is organized as follows: In Section II. Related works are discussed. Section III presents Methodology. Experimental analysis and results are described in Section IV and conclusion and future work is discussed in Section V.

II. RELATED WORKS

This section discusses related works in the specific area of twitter sentiment analysis.

Masashi et al [1] proposed aspect identification method for analyzing sentiments in review documents. They applied non-tagged data and clustering approach to solve the problem of the number of training data classifying similar sentences into clusters first then the aspects of sentences that are close to the centroid of each cluster were tagged. They identified the aspect of sentences in test data using SVM with 73.9%.

Shahana et al [2] presented selected features from high dimensionality of feature set using feature selection techniques such as information gain, TF-idf, Chi-square and mutual information. The methods were evaluated over movie review dataset from websites. The performance was evaluated using SVM and Weka tool. They proved that unigram using stemming with stop words give high accuracy.

Deepa et al [3] performed aspect based sentiment analysis on movie reviews. The aspect as well as sentiment detection using clustering, review guided clustering and manual labeling

was evaluated against manually constructed test data set. The experimental results showed better performance.

Li et al [4] analyzed the usefulness of labeled data from disaster along with unlabeled data to learn domain adaptation classifiers for their target. The results showed that domain adaptation approaches with target unlabeled data other than labeled data are superior.

Rishab et al [5] proposed a cluster then predict approach to analyze the sentiments of a launched product, the results proved an efficiency of 74% with the hybrid approach.

Keng et al [6] presented a sentiment topic recognition model based on correlated topics with the proposed method called "Variational Expectation Maximization algorithm". The results showed that lexicon based approach used in sentiment detection could detect basic sentiments, but inadequate in detecting figurative expression.

Ghazeleh et al [7] proposed a lexicon based model to analyze sentiments and its applications in disaster relief. The studies proved that more complex machine learning technique along with strong features were required for better accuracy and also to easily track the changes during disasters and to make quick decisions, visualization techniques should be improved so as to allow real time visual analytics of disaster related posts.

Swapna et al [8] discussed various techniques and approaches of twitter sentiment analysis of public and made a comparative study of existing papers.

Neha et al [9] gives an overview of existing methods and the recent advances in the area of sentiment analysis and have provided a layout for the future directions in this field.

Gayatri et al [10] proposed a lexicon approach (LDA), foreground and background LDA in order to filter the foreground topics and filter out background topics. They also analyzed sentiments on twitter data using pos tagging and also used multi-core programming so as to perform parallel processing in multi-core.

G.Sneha et al [11] provided the survey regarding the challenges and overview of some of the clustering and classification algorithmic techniques used for sentiment analysis and opinion mining.

Thomas et al [12] investigated the method to optimize dualist architecture for agile sentiment analysis. They performed analysis with several semi supervised learning algorithms with Naïve Bayes model and showed how the modifications could improve the performance of bespoke classifiers on large datasets.

Piyoros et al [13] presented a model for the prediction of Sentiment scores for sarcastic and ironic tweets. The results showed that the proposed method out performs the existing methods on the related tasks.

Walid et al [14] proposed and evaluated a filtering approach regarding six broad topics and were tested on four different time periods over four months time. The results showed that the proposed approach achieved 84% increase in recall for the baseline- approach.

Korkmaz et al [15] proposed a two stage semi supervised model for sentiment classification based on priority aging and expectation maximization algorithms. The results showed better performance in the classification of sentiments.

Xiuzhen et al [16] proposed an effective approach to incorporate the knowledge of word labels into expectation maximization process for sentiment classification of documents. The results showed that a combination of limited domain specific labeled training documents with general lexical knowledge could achieve better performance.

Mustafizur et al [17] proposed a hidden topic sentiment model to capture topic coherence by tracking changes in sentiments and utilized the linguistic cues to guide topic and sentiment transitions. The experiments shows pretty good quality evaluation results for interleaved documents.

Arti et al [18] proposed an approach for online analysis using EM algorithm. The approach recognizes the technical feature value depending on the reviews that were summarized.

Mathew et al [19] proposed a body of classified customer postings from companies and also developed a sentiment analysis matrix from those classifications.

Gangli et al [20] presented new techniques to extend the capability of cluster based sentiment analysis approaches. Experimental results proved that the cluster based approaches gives high quality sentiment analysis result and are suitable for recognizing neutral opinions.

A. Discussion

This section discusses the summary of the above mentioned related works of recent 4 years.

The paper [1] discusses the aspect identification of sentiments. Bayon tool was used for analysis. The data set used was game review documents and was collected from the website "http://ndsmk2.net/. The Bisection and Support Vector Machine algorithms were the methods used for analysis and accuracy as the metric for evaluation. In paper [2], the researchers performed the evaluation of features on Sentiment analysis using Weka tool and the datasets collected from website but the details are not specified. Accuracy was used as the metric for evaluation. In Paper [3], the researchers have performed a semi supervised aspect based sentiment analysis for the movies using "review filtering technique". The

movie review dataset collected from Amazon, movie /TV shows. The performance was evaluated based on the metrics score and semantic similarity using the algorithm K Medoid. In paper [4], researchers used twitter information for detecting the diaster response on dataset collected based on disastrous Hurricane study and Boston marathon bombings. Naive Bayes algorithm was used for performing the experiments. The metric and tool used for evaluating and conducting the experiments were not clearly mentioned. More number of datasets is required to come up with more general conclusion. In paper [5], a hybrid approach including K means and CART algorithm was used for analyzing sentiments of a launched product. Tool used was not specified. Tweets were collected for the analysis but the source of data was not clearly mentioned. The accuracy was taken as the metric for evaluation. The researchers in paper [6] proposed a lexicon based approach to analyze sentiments of airline quality rating. The tweets were collected based on Air Tran Airways, Frontier and Skywest airlines. The metric used was Air Quality Rating Score and R tool was used for the analysis. The approach was inadequate in detecting figurative expressions. In paper [7], an overview of sentiment analysis in social media and its applications in disaster relief were discussed. IN-SPIRE and VISA were the tools used for the experiments. The information of datasets was not mentioned clearly. The Visualization frequency was used as the evaluation metric. In paper [8], researchers performed a review on sentiment analysis of the twitter data. Clarity is missing. The researchers in paper [9] used the Maximum likelihood probability and TF-IDF as the metric for analysis. Other details were not mentioned. In paper [10], the algorithm used was Latent Dirichlet Allocation in a lexicon based approach. The datasets used were the twitter information but the source was not clearly mentioned. The Weight and POS tagging was used as the metric for conducting the experiments. The paper [11] , used K Means and Self Organized Map (SOP) as algorithms and techniques as supervised and unsupervised. Other details regarding the dataset, source, metric and how to perform the analysis was missing. In paper [12], the analysis was performed with Naïve Bayes algorithm and the technique was Natural Language Processing (NLP). The datasets used were 24 Twitter datasets and were collected by social scientists and from the existing datasets of Norrie2015. The metric used was F1 score, not mentioning the tool used for the experiment. Paper [13] proposed a sentiment analyzer for ironic and sarcastic data using Weka tool. The details of datasets were not mentioned clearly. It was collected from those released by the organizers of the SemEval 2015, task11. The metrics used were the Cosine Similarity and Root mean squared error (RMSE) and a supervised technique, REP tree algorithm was used for the analysis. In paper [14], the datasets used were Arabic tweets and was collected from Tweet Mozag. The Filtering and baseline approach and Support vector Machine algorithm were used for identifying broad dynamic topics. Researchers in paper [15] used were movie review datasets and were collected from IMDB. The technique used was supervised and EM and Naïve Bayes were the algorithms. The F measure and Recall values were used as the metric for performance evaluation. The tool used was not specified. The researchers [16] used was cartoon, Mc Donald datasets and were collected from TREC Blog06 collection. Lexicon EM algorithm and semi supervised algorithm were used for analyzing sentiments. Metric were collected from TREC Blog06 collection. Lexicon EM algorithm and semi supervised algorithm were used for analyzing sentiments. Metric used was Accuracy. The tool used for the analysis was not specified. In paper [17], Naïve Bayes algorithm was used and the technique was semi supervised. The datasets used were the product reviews of four categories TV, tablet, camera and phone. It was collected from Amazon and New Egg. MS^m was used as the metric for evaluation. Tool used was not mentioned in the paper. In paper [18], Technique used was NLP. The dataset and data source details are not provided. Precision was used as the metric for evaluation. In paper [19], the datasets used were subject line and posting text and were collected from Amazon, HP, and Apple. Decision tree and Discriminant Analysis were the algorithms used. The techniques used tools and metric for evaluation were not specified. In paper [20], researchers presented a new technique to enhance the capability of cluster based sentiment analysis approaches. Accuracy was measured as the metric for evaluation. The review dataset was collected from website but the other details were not mentioned.

In our proposed work we have presented a novel unsupervised fuzzy clustering method to analyze the brand impact. The real datasets of 300 samples of tweets regarding the brand Samsung Galaxy S6 was collected from Twitter API gathered over a period of 1 year was used as the data set. Then a comparative study has been made with the existing K means and EM algorithms using the evaluation metrics Accuracy, Precision, Recall and Execution Time. It shows an overall accuracy of 76%.

III. METHODOLOGY

The following conceptual diagram illustrates the steps involved in analyzing twitter data for our sentiment analysis.

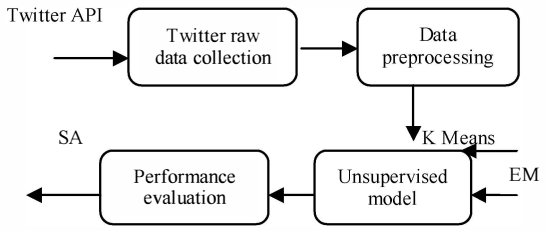

Fig 1: Conceptual diagram of methodology

A. Twitter Data Collection

We have collected 300 samples of real datasets from twitter API regarding the opinion of a brand namely Samsung Galaxy S6 to analyze and to predict the brand impact. The real twitter datasets were collected over a period of one year (2015-16). The R tool version 3.2.1 was used for extracting tweets from twitter API. The raw datasets collected are then employed to a filtering process called Data Preprocessing. It is described in the section below.

B. Data PreprocessingTwitter

After collecting the raw data of tweets, it is further sorted and collated using various filtering tasks such as Preprocessing of Tweets. It involves stages such as Replacing URL, Removal of stop words, Pointer detection, Punctuation Identification, Removal of irrelevant words and Compression of words. Replacing URL replaces URL with an equivalent keyword <URL>. Removal of stop words involves replacement of emoticons with corresponding keywords. NLP package was used for the process of identifying and removing stop words. The Pointer detection stage abstracts hash tags and user names with equivalent keywords <USER> and <HASHTAG>. Punctuation Identification process removes irrelevant punctuations within the tweets with a corresponding keyword<PUNCT>. Removal of irrelevant words involves the act of removing meaningless and obscure words and Compression of words compresses elongated words that are used to express strong emotions. For eg: wooooow, cooool etc.

C. Sentiment Analysis using Unsupervised K Mean Clustering Method.

Simple K-means clustering presents a formal definition as an optimization problem such as: The k- cluster centers to be obtained and assign the objects to the nearest cluster center as the squared distances from the cluster are minimized. The optimization problem itself is known to be NP- hard, and hence the common approach would be to search only for the approximate solutions. A particularly well known approximate method is Lloyd's algorithm, [21] often referred as "k-means algorithm".

The pseudo code for performing sentiment analysis with K-Means is described below:

Algorithm:

Input: Dataset D=d_1, d_2, ……. d_k

Number of clusters (K)

Steps:

1. Initialize the value of K
2. Centroids are chosen by selecting K random points of the datasets
3. Assign data objects to their nearest cluster centroid based on the Euclidean distance measure.
4. Determine the mean or centroid of all data objects in each cluster.
5. New centroid positions are updated
6. Repeat steps 2,3,4,5 until convergence.
   Output: (i) Cluster centroids (ii) Cluster labels of D

The K Means algorithm was used for finding the hidden patterns of our unlabelled twitter dataset of 300 samples regarding the opinion of a brand namely Samsung Galaxy S6. Here we have initialized the value of K as 2. The objects were randomly chosen as the centroids. Based on the Euclidean distance measure the objects were assigned to the closest centroid. Then the mean of each cluster was calculated and the new centroid positions were updated and the processes were repeated till the same points were consecutively assigned to each cluster. Here the three clusters obtained categories Positive, Negative opinions of the brand so as to predict the brand impact.

D. Sentiment Analysis using Unsupervised EM Clustering Method

The EM algorithm was explained and given its name in a paper, classic 1977 by Arthur Dempster, Nan Laird, and Donald Rubin [22]. The EM algorithm is used to obtain maximum likelihood parameters of a statistical model in the cases where the equations cannot be solved directly. Typically there were missing values within the data or the model can be developed more simply by assuming the existence of additional unobserved data points.

The pseudo code for performing sentiment analysis with EM is described below:

Algorithm:

Input: Dataset D=d_1, d_2, ……. d_k
Total number of clusters (K)
Accepted error to converge

Steps:

1. Iterate through Expectation step that estimate the probability of each point of the cluster.
2. Iterate through Maximization step, estimates the parameter vector of the probability distribution of each class.
3. Repeat the step 1, 2 until the distribution parameters converges**.**
4. Output: Parameters of probability distribution with maximum likelihood value of the attribute.

The method was applied on the twitter datasets of 300 samples regarding the opinions or sentiments of a brand namely Samsung Galaxy S6. The method iterates through two process expectation and Maximization process. The first process estimates the probability of each point of the cluster. Then the second step, which is the Maximization step that estimates the parameter vector of the probability distribution of each class where the classes assigned in our work, was a three class problem Positive, Negative and Neutral. The two steps where repeated until it reaches the maximum number of iterations. Here in our experiment we have assigned the number of clusters K as 2. The Clusters obtained where Cluster 0 as Positive, Cluster 1 as Negative.

E. Sentiment Analysis using Proposed Fuzzy Clustering Method

Algorithm:

Input: Dataset D=d_1, d_2, ……. d_k
Total number of clusters (K)

Steps:

1. For each object j and each cluster c there will be a membership m_{jc} . It satisfies the below conditions:

(i) $m_{jc} \geq 0$ for all j=1,….n and c=1,….k

(ii) $\sum_{c=1}^{k} m_{jc} = 1 = 100\%$ for all j=1,…n

2. Minimize the objective objection
3. Compute Fuzziness coefficients such as Dunn's partition coefficient and normalized.
4. Compute average Silhouette width per cluster.

Output:

(i) Membership of all objects

(ii) Nearest Crisp Clustering

The proposed method is a modified fuzzy clustering algorithm. The method incorporates the functioning of PAM partition based algorithm as well as FANNY clustering algorithm. In this method, each object is dispersed widely over various clusters and the degree of belong to an object to unlike clusters is quantified by means measurement called membership coefficients. The membership coefficients normally range from 0 to 1, with the condition that the sum of their values is one. This concept is called Fuzzification of the cluster configuration. The proposed method has an advantage that it does not force every object into a particular cluster. The method was employed to 300 twitter samples of real data sets. The objective function obtained was1818.394 and a maximum of 12 iterations were occurred. To determine how hard or fuzzy the clustering is, two fuzziness coefficients were computed. They were Dunn's partition coefficient and Normalized. Dunn's coefficient value obtained was 0.8243 and the Normalized coefficient was0.6487. Average Silhouette widths per clusters obtained were 0.7981 and 0.7966. The method specifies how accurate the percentages of opinions regarding the brand Samsung galaxy such as Positive, Negative and Neutral sentiments are to predict the brand impact.

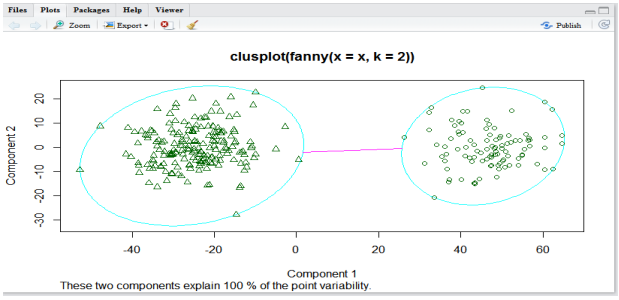

Figure 2: Clus plot of the proposed method

The Clus plot of the proposed method is illustrated in the figure 2 shown below. Here the Component2 was represented in the Y axis of the plot and Component 1 was plotted in X axis. The number of cluster selected was 2 .These components explained 100% of the point variability.

IV. Experimental Analysis

This section discusses about the performance analysis of proposed method and the comparative analysis of other two common partition based clustering methods namely K Means clustering and EM clustering method. The methods were employed with the collected samples of real datasets containing 300 samples of twitter information regarding the sentiments of a particular brand namely Samsung Galaxy S6, with the purpose of choosing the efficient clustering model for finding percentage of the correctly clustered instances and to predict the brand impact.

The experiment was carried out using an open source data mining tool R 3.2.1 version on Pentium R processor with memory of 2GB RAM.. The metrics used here for evaluating the performance of the clusters are accuracy, precision, recall and the execution time.

Definition 1: Accuracy

The accuracy is defined as the percentage of correctly clustered instances, where Tp is the true positive, Fn is false negative and Fp is false positive.

$$Accuracy = \frac{Tp+Tn}{Tp+Tn+Fp+Fn} \quad (1)$$

Definition 2: Precision

Precision is the fraction of retrieved documents that are relevant to the query:

$$Precision = \frac{Tp}{Tp+Fp} \quad (2)$$

Definition 3: Recall

Recall is the fraction of the documents that are relevant to the query that are successfully retrieved.

$$Recall = \frac{Tp}{Tp+Fn} \quad (3)$$

The accuracy or the percentage of correctly clustered instances obtained with K Means clustering technique on twitter information is 75.5 % where as EM produces 63.4% accuracy. However the proposed method outperforms these two methods with an overall accuracy of 76.4%. The precision and recall measures obtained by the proposed method are 0.57 and 0.33 where as K Means produced 0.55 Precision and 0.36 Recall value. EM produced Precision as 0.43 and 0.33 Recall value. The metric representation is demonstrated below in table 3:

Table 3: The Metric representation for Twitter SA

Accuracy	Time	Techniques	Precision	Recall
75.5%	0.25	KMeans	0.55	0.36
63.4%	1.06	E M	0.43	0.33
76.4%	0.18	Proposed Method	0.57	0.33

The execution time required for building the K means

clustering technique in the domain of twitter sentiment-analysis is 0.25 seconds, EM Partitioning clustering technique required 1.06 seconds to build up the model. The proposed method required only 0.18 seconds which is comparatively low compared to the other two clustering models. The figure 3 shows the execution time of methods.

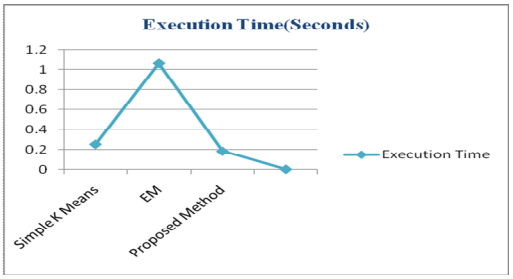

Fig3: Execution time required by the partition based methods

The metric used for evaluating the performance of the partition based clustering methods are Accuracy, Recall and Precision. The Accuracy, Precision and Recall values obtained by K Means, EM and the proposed method are represented in the graph below:

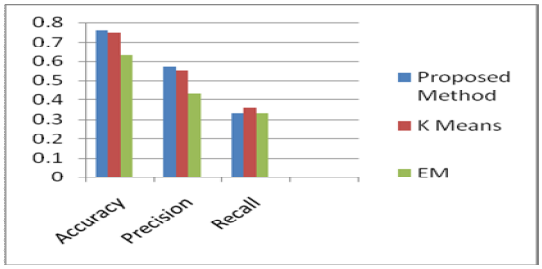

Fig 4: Metric representations of the methods

The brand impact of the trendy product Samsung Galaxy S6 predicted using the proposed method was 60% Positive, 30% Neutral and 10% Negative Sentiments or opinion.

V. Conclusion and Future Work

The Partition based clustering techniques would provide accurate results without manual processing, linguistic knowledge or training. As per the experimental analysis, the proposed approach is proven to be useful in performing high quality results in the domain of twitter sentiment analysis. The Simple K means Partitioning clustering technique is more efficient than Expectation-Maximization (EM) Partitioning clustering technique with an overall accuracy of around 75.5% and time taken to build the model requires 0.1 seconds as against EM with 1.06 seconds. However the proposed model gives a pretty good accuracy of 76.4 and a required less time to build the model compared to the other two methods. The model is proven to be more suitable for real time applications on the aspect of accuracy and execution time. As the future work, with more samples of twitter feeds , the performance of Partitioning clustering techniques would be analyzed against other clustering techniques and would come up with another novel approach to analyze twitter sentiments.

References

[1] W.Magdy,Tamer Elsayed, "Unsupervised adaptive microblog filtering for broad dynamic topics," Elsevier, vol. 52, pp. 513-528, 2016.

[2] H.P Shahana, Bini Omman, "Evaluation of Features on Sentiment Analysis", Elevier, International Conference on Information and Communication Technologies, vol. 46, pp.585-592,2015.

[3] Deepa Anand, Deepan Naorem, "Semi supervised Aspect Based Sentiment Analysis for Movies using Review Filtering", Procedia Computer Science, vol.84, pp. 86-93, 2016.

[4] Hongmin Li, Nic Herndon, K.Neppalli, "Twitter mining for Disaster Response:A Domain Adaptation Approach",Proceedings of the ISCRAM Conference, pp. 24-27, 2015.

[5] R.Soni, K.James Mathai, "Effective Sentiment Analysis of a Launched Product using Clustering and Decision trees",International Journal of Innovative Research in Computer and Communication Engineering ,vol. 4, pp, 2016.

[6] M.North,S.Riniker, "Consumer sentiment extraction from unstructured data",Issues in Information Systems,vol. 15, pp. 430-433, 2014.

[7] Gang Li, Fei Lu, "Sentiment analysis based on clustering:a frameworkin improving accuracy and recognizing neutral opinions,"Springer Science+ Business Media Newyork, 2013.

[8] Esi Adeborna, Keng Siau, "An Approach to Sentiment Analysis-The Case of Airline Quality Rating", 2015.

[9] G.Beigi,Xia Hu,R.Maciejewski, Huan Liu, "An Overview of Sentiment Analysis in Social Media and its Application in Disaster Relief," Elsevier, 2014.

[10] S.R Kharche,L.Bijole, "Review on Sentiment Analysis of Twitter Data," International Journal of Computer Science and Applications,vol.8, 2015.

[11] N.S Joshi,S.A.Itkat, "A Survey on Feature Level Sentiment Analysis ," International Journal of Computer Sciencence and Information Technologies,vol. 5(4), pp. 5422-5425, 2014.

[12] G.S Potdar,R.N.Phursule, "International Journal of Science and Research," vol. 4, pp. 2319-7064, 2015.

[13] G.R Bhongade,Pragati Patil, "A Novel Approach for Analyzing the Public Sentiment Variations on Twitter Using Multi-Core Programming ," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 5, 2015.

[14] G.Sneha,C.T VidhyaW, "Algorithms for Opinion Mining and Sentiment Analysis," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 6, 2016.

[15] T.Kober,D.Weir, "Optimising Agile Social Media Analysis," 2015.

[16] P.Tungthamthiti, E.Santus,H.Xu, Chu-Ren Huang, K.Shirai, "Sentiment Analyzer with Rich Features for Ironic and Sarcastic Tweets," 29th Pacific Asia Conference on Language ,Information and Computation, pp. 178-187, 2015.

[17] Arti Buche,M.B Chandak,A.Zadgaonkar, "An Approach for Online Analysis using Expectation Maximization," International Journal of Innovative Research in Computer & Communication Engineering, vol. 1, 2013.

[18] Y.Korkmaz Yengi, M.Karayel, S.I Omurka, "An Alternative Method for Sentiment Classificatio with ExpectationMaximization & Priority Aging," International Journal of Scientific Research in Information Systems & Engineering, vol. 1, 2015.

[19] Md Mustafizur Rahman,H.Wang, "Hidden Topic Sentiment Model," pp. 155-165, 2014

[20] X.Zhang, Y.Zhou,J.Bailey,K.Rama Mohanarao, "Sentiment Analysis by Augmenting Expectation Maximization with Lexical Knowledge",2013.

[21] Lloyd, S, "Least squares quantization in PCM", IEEE Transactions on Information Theory, 28(2), pp.129-137,1982.

[22] Nan Li, D.D Wu. "Using text mining and sentiment analysis for online forums hotspot detection and forcast",Elsevier, pp.354-368,2010.