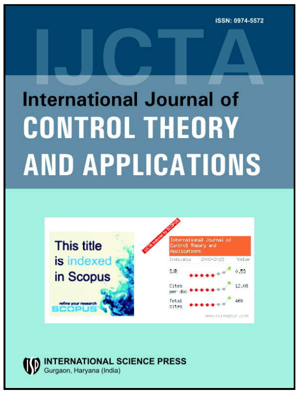

International Journal of Control Theory and Applications

ISSN : 0974-5572

© International Science Press

Volume 10 • Number 19 • 2017

A Novel Cluster Based Unsupervised Technique for Twitter Sentiment Analysis

Hima Suresh[1] and Gladston Raj S.[2]

[1] *School of Computer Sciences, Mahatma Gandhi University, Kottayam, Kerala, India*
[2] *Department of Computer Science, Government College, Trivandrum, Kerala, India*

Abstract: Sentiment Analysis is one of the most promising research areas in data mining discipline. It actually centers on analyzing opinions relating a particular subject of interest using different Sentiment Analysis approaches, predominantly Machine Learning Approach. Micro blogs especially Twitter has become one of the fastest and user friendly platforms for online information sharing with the explosive growth of online social media networks. We have combined the Sentiment Analysis concept with a modern spin, the mining of Twitter feeds. In this paper, we introduce the Fastest Threshold Clustering Algorithm in the domain of Twitter Sentiment Analysis, which was never tried by any researchers earlier in this specific area and also a novel method is designed and implemented to analyze the impact of a trendy product using the real data samples gathered from the micro blog; Twitter taken over a duration of one year. Experimental result shows that our method statistically outperformed the commonly used traditional Unsupervised clustering techniques in terms of Accuracy.

Keywords: Sentiment Analysis, Clustering, Unsupervised Learning

1. INTRODUCTION

Twitter has emerged as a new medium for content sharing and information gathering. It is increasingly being used by many social institutions as well as government organizations to gather responses from targeted audience. Traditionally most of the researches have focused on assorting larger texts such as reviews. Tweets are entirely different from reviews primarily because of their purpose; Twitter feeds are limited to 140 character text and are more casual whereas the reviews represent summarized ideas or thoughts of the users.

Twitter services in recent times has become one of the popular social networking sites among internet users to generate millions of tweets expressing the sentiments regarding brands or other products, to post, repost and to discuss opinion of a variety of trendy topics. The analysis of Twitter feeds in this regard, using a Sentiment Analysis approach would be essential in gathering cyclical patterns from the social behavior.

Sentiment Analysis refers to the computational identification, extraction and categorization of the subjective information expressed in a text to determine the opinion or attitude of a writer about the topic. Sentiment

Analysis is otherwise known as opinion mining. The opinions are not necessarily based on fact or knowledge; it could be the views of a user regarding a product or it might be the reviews about current topic.

Sentiment Analysis approach is mainly categorized into two; Supervised and Unsupervised Learning Techniques. In this paper, we adopted the Unsupervised Learning Technique namely the Fastest Threshold Clustering Algorithm and a modified method is proposed and implemented using the real 1500 Twitter samples of data sets. The purpose of designing this method is to analyze the patterns of tweets which are further sorted into three class sentiment values such as positive sentiments, negative sentiments and neutral sentiments.

We summarize the main contribution of our work as the following:

(1) The real data samples of 1500 Twitter feeds were gathered from the Application Programming Interface of Twitter taken over a time period of 1 year.

(2) We have employed a novel idea of Fastest Threshold Clustering concept in the discipline of Twitter Sentiment Analysis.

(3) Made a comparative analysis with the commonly used Unsupervised clustering techniques.

(4) Designed and implemented a novel method to attain higher efficiency in analyzing the product brand impact of the smart phone “Samsung Galaxy S6”.

The remaining section of the article is organized as : In section 2, we have made a detailed study of the recent research related works. The methodology of our work has been discussed in section 3. Section 4 describes the details of empirical analysis and Section 5 presents conclusion.

2. RELATED WORK

In this section, a brief overview of one of the common Sentiment Analysis approaches namely the Unsupervised Learning Technique was presented and a study of related works for the past 4 years (2013-2016) was also conducted to identify the shortcomings in Twitter Sentiment Analysis using Unsupervised Clustering Technique.

2.1. Unsupervised Learning Technique

An Unsupervised Learning Technique is a kind of Sentimental Analysis approach coming under the Machine Learning Technique that is used to draw illations from the datasets which comprises input data without any labeled responses. One of the most common Unsupervised Learning Techniques is the clustering method. The Clustering is defined as the task of grouping similar set of objects in such a way that objects belonging to the same group should be more alike to each other than to those in other groups.

Apeksha et al [1] presented a review on the existing document clustering techniques as well as cluster based classification techniques. They have also discussed the preprocessing tasks done on text classification.

Gang Li et al [2] introduced novel techniques to extend the capability of cluster based Sentiment Analysis approaches. The result shows that the adopted method was desirable for recognizing data containing neutral opinions while performing Sentiment Analysis on web documents. Researchers used a modified mechanism for voting as well as a distance measure approach.

Bjorn et al [3] applied clustering technique for Sentiment Analysis to detect and code new frames. The research has several limitations that they have used only 3 newspapers from 2 different countries as the data source. No true reference lists of frames were used in the content of nuclear power debate. Each time an analysis was conducted and was observed that the result varies accordingly, since the work was analyzed using a non-deterministic technique.

Li Zhao et al [4] presented an Unsupervised model in Posterior Regularization Framework in order to cluster the aspect related phrases. The model was proposed for performing Aspect Level Sentiment Analysis.

Nihalahmad et al [5] presented a method to classify movie reviews based on the analysis and extraction of appraisal groups like comedy, thrill, action etc. The proposed JIBCA system was evaluated on existing IMDB movie reviews and other blog documents. The work shows 60% accuracy results.

Linhong et al [6] proposed a tri-clustering frame work, that is an Unsupervised approach. This framework analyses sentiments through the co-clustering process of a tripartite graph. The analysis shows pretty good results.

Souniel part [7] presented an approach to deal with the expression of sentiment patterns captured from the tasks of TASS 2015. The experimental result achieved 0.63 and 0.56 accuracy for Social TV and STOMPOL corpus. The work could be further improved by incorporating it with more language processing methods.

Leonardo et al [8] proposed a model for clustering time series data through community detection in complex networks. The researchers used a naive method for community detection. The work could be improved by using an automated strategy for selecting the relevant number of neighbours and to speed up the required network construction methods.

Ashish et al [9] proposed a modified Unsupervised Clustering Technique incorporating K Means and Naïve Bayes algorithms. The researchers used mobile review datasets collected from Amazon and Flipkart to perform the analysis.

Luiz et al [10] presented a model with SVM classifier combined with a cluster ensemble to show that the efficiency in classification would be better using the hybrid approach than using the Supervised Learning method ; SVM alone. The existing Sanford dataset was used for the analysis of Tweets.

Pooranam et al [11] proposed a method to extract opinion from stock data. Researchers have developed a subspace clustering method for outlier detection. The data source and the details of the datasets used were not clearly mentioned in the paper.

Ayesha et al [12] presented an Unsupervised Machine Learning Technique ; Spectral Clustering method to group the Tweets into positive and negative clusters. The researchers used existing movie review dataset but the data source was not specified clearly. The experimental result shows 70% accuracy with the Spectral Clustering method in this scenario.

Abeed et al [13] presented a Supervised sentiment classification model that was competed in the existing Sem Eval 2016 Task 4. The model achieved 0.694 and 0.650 F scores for positive class and 0.391, 0.493 for negative class.

3. METHODOLOGY

3.1. Twitter Data Acquisition

The real Twitter data samples of 1500 datasets were gathered from Twitter Application Programming Interface (API) using a Free software called R, regarding the opinion of a trendy product brand “Samsung Galaxy S6” to both examine and to predict the product brand impact with a proposed Unsupervised Clustering method. The real datasets were gathered from Twitter over a time period of one year from 2015-16. The data mining tool R for Windows 3.2.1 version was used for extracting the tweet information from Twitter API. The raw datasets collected were sorted based on a filtering task known as the Twitter data Preprocessing task. It is described in the following section 3.2 .

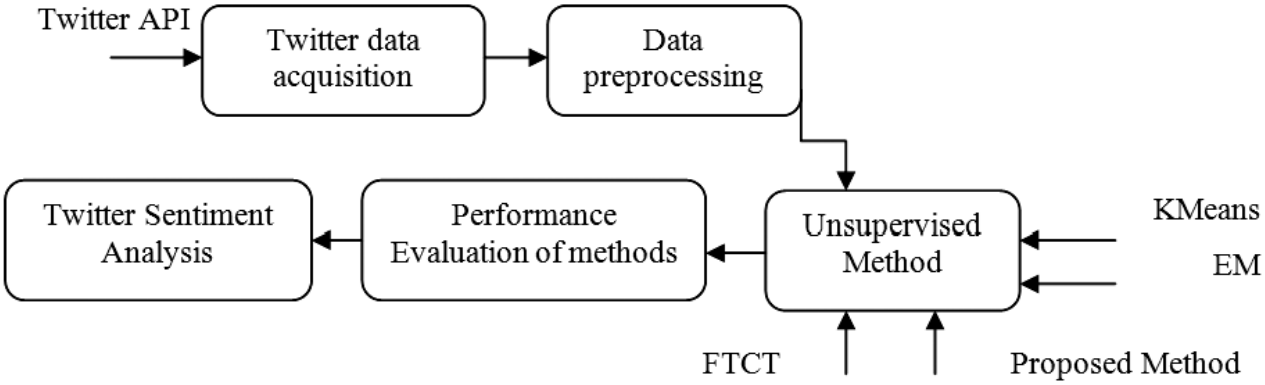

Figure 1: Architectural diagram of the proposed model

3.2. Twitter Data Preprocessing

After the acquisition of the raw Twitter data sets, it is then applied to Twitter Data Preprocessing. It is categorized as Data Cleaning and Normalization. Data cleaning and normalization involves various stages namely the URL Replacement , Stop words Removal, Detection of Pointers, Detection of Punctuation, Discarding irrelevant words , Stemming and the Word Compression.

(1) URL Link Replacement:-The URL Link Replacement stage identifies all the hyperlinks in the tweets and replaces the same with an alternate keyword URL.

(2) Removal of irrelevant Stop Words: - The Stop Words are commonly used functional words in English that does not carry any specific meaning. Eg: 'or', 'of', 'and' etc. The Removal of irrelevant Stop words stage involves the process of replacement of words that doesn't contain any sentiments or emotions. In order to identify and remove stop words from the collected Twitter feeds we have used the Natural Language Processing package in our work.

(3) Detection of Pointers:-The stage "Detection of Pointers" identifies hash tags "#" and user names "@" from each tweet samples with an equivalent keyword namely USER and HASHTAG.

(4) Detection of Punctuation:-The Detection of Punctuation stage identifies and discards all the punctuations that are found irrelevant in the tweet samples with an alternative keyword namely PUNCT.

(5) Discarding irrelevant words: - This stage "Discarding irrelevant words" involves the process of eliminating all the unclear and meaningless words in the tweet sample.

(6) Process of Stemming:-The Stemming stage reduces derived words to their corresponding stem. It allows almost similar terms into a single stem. Snowball; an open source library in data mining tool R has been used here.

(7) Word Compression: - The Word Compression stage involves the act of compressing lengthy words that express strong emotions in tweets. If the occurrence of such word is repeated then it would be limited to two occurrences. For eg: Coooool, Wooooow etc.

3.3. K Means Clustering method for performing Sentiment Analysis

K-means clustering method is a vector quantization method that finds the cluster centers and the objects will be assigned to the closest cluster centroid based on the minimum squared distance obtained from the cluster . The standard K means algorithm was first proposed by Lloyd [14].

The pseudo code for Twitter Sentiment Analysis using K-Means clustering method is shown below:

Pseudo code:

```
Input:  Twitter dataset $D = \{d_1, d_2\ldots, d_k\}$, Set of centers $C = \{c_1, c2, c3\ldots\ldots, c_k\}$, Number of clusters
Steps:
(1)  Assign the number of clusters.
(2)  Randomly select the C cluster centroids.
(3)  Assign data points to their closest cluster centroid based on the Euclidean distance measure.
(4)  Calculate the centroid of all data points in each cluster.
(5)  Newly obtained centroids are updated.
(6)  Repeat the above mentioned steps 2, 3, 4 and 5 until convergence.
Output Obtained: (i) Cluster centroids; C (ii) Cluster labels of Twitter dataset
```

The Simple K Means clustering method was used for detecting hidden patterns in our unlabelled data samples of 1500 real datasets about the sentiments of a popular brand "Samsung Galaxy S6". Here we have initialized 3 as the number of clusters for performing the analysis. The data points were randomly chosen as the centroids C. Based on the distance measure called Euclidean distance , data points were assigned to the nearest cluster centroid. The mean value of the cluster was estimated and the newly obtained centroids were updated accordingly. The steps were repeated until the similar data points were consecutively allocated to each cluster. The clusters generated was categorized into Positive, Negative, Neutral sentiments of the popular brand to predict the impact of the corresponding product brand.

3.4. Unsupervised EM Clustering method for Sentiment Analysis

The Expectation Maximization clustering method was first described in a classic 1977 paper and was given its name by Arthur Dempster, Nan Laird, and Donald Rubin [15]. The EM clustering algorithm is said to be an iterative aspect that is used to get the maximum likelihood or maximum posteriori parametric quantities of a statistical model in the cases where equations cannot be figured out right away. Generally the model involves missing values among the data The model can be devised more simply by presuming the existence of additional unseen data values.

The detailed pseudo code for performing Twitter Sentiment Analysis using the EM Clustering method is shown below:

Pseudo code:

```
Input:  Twitter dataset $D = \{d_1, d_2\ldots, d_k\}$, Number of clusters to be assigned, The Accepted error to converge
Steps:
(1)  Repeat through first process called the "Expectation" that estimates probability of each data point in
     the cluster.
(2)  Repeat through second process called the "Maximization" that estimates the parameter vector of
     probability distribution of each category.
(3)  Repeat the steps 1 and 2 until the distribution parameters reaches a stable end point.
Output Obtained: Probability distribution parameters with maximum likelihood.
```

The EM clustering method was employed to our Twitter feeds of 1500 real data samples to perform Sentiment Analysis. The algorithm generally repeats through two processes namely Expectation and Maximization process. The Expectation process calculates the probability of each data point of the cluster and the Maximization process calculates parameter vector of the probability distribution of each category of the class. The processes would be iterated until it accomplishes maximum number of iterations. The number of clusters assigned in our case is 3 where the cluster labeled 0 represents Positive sentiments, the cluster labeled 1 represents Negative sentiments and finally the cluster labeled 2 represents Neutral sentiments.

3.5. Proposed Method for Twitter Sentiment Analysis

The Fastest Threshold Clustering Technique was designed and developed by David Varadi. The method works almost similar to the Minimum Correlation Algorithm. It determines how close or far away an asset related to the chosen asset by using the average correlation of each asset to other assets.

The proposed model is ‘Modified Fastest Threshold Clustering Technique (MFTCT)’. The model was applied to our real 1500 Twitter data samples to predict the brand impact. We have used a correlation threshold of 0.5 to separate the similar assets from dissimilar ones. This method has some desirable properties that the traditional clustering methods does not possess such as (a) More stable clusters were produced (b) Method is faster and deterministic. The accuracy obtained by the proposed method is higher than the existing algorithms. The proposed method predicted 65% positive, 20% neutral and 15% negative sentiments for the brand Samsung Galaxy S6.

The detailed pseudo code for Twitter Sentiment Analysis using the proposed method is shown below:

```
Pseudo code:
Input:  Twitter dataset, $D = \{d_1, d_2\ldots\ldots dk\}$, The minimum count of vectors
        Distance threshold for each cluster threshold
Steps:
(1) Assign the assets to a cluster
    (a) If one asset is remaining then form a new cluster
(2) Determine the highest average correlation to all assets not yet been assigned to any cluster.
(3) Determine the lowest average correlation to all the assets not yet been assigned to a cluster.
(4) If both the correlations are greater than threshold
    (a) Form a new cluster with highest average correlation and lowest average correlation
    (b) Add to the new cluster all the assets not yet assigned to a cluster
    (c) Get the average correlation to highest and lowest average correlation greater than threshold
(5) Otherwise, form a cluster made of higher average correlation
    (a) Add to the new cluster, all non assigned assets.
        (i) Get the correlation to higher average correlation greater than threshold
    (b) Form a cluster made of lowest average correlation
        (i) Add to the cluster all non assigned assets.
        (ii) Get the correlation to higher average correlation greater than the threshold
(6) Calculate the accuracy, precision and recall measures
(7) Resample the results with the parameters.
Output: Cluster labels of Dataset D
```

4. EMPIRICAL ANALYSIS

This section describes the experimental results of our proposed method for Sentiment Analysis compared to the other commonly used Unsupervised clustering algorithms; K Means clustering, Expectation Maximization clustering (EM) and Fastest Threshold Clustering Technique (FTCT). The Real Twitter data samples of 1500 dataset gathered over a period of 1 year, were applied to these methods with the aim of selecting the most efficient Unsupervised clustering model to predict the brand impact of a popular brand "Samsung Galaxy S6". The experiment was executed successfully using free data mining software called R of version 3.3.2 on Intel R core processor with the 2GB memory RAM. In order to perform the analysis, we have chosen three different metrics as to compare the clustering methods in this domain.

The metrics used in our research work for executing the performance analysis of the commonly used clustering methods are accuracy, precision, recall.

Definition 1: Accuracy

Accuracy metric is defined as the percentage of correctly clustered events of tweets. The formula for finding the accuracy measure is shown below:

$$Accuracy = \frac{Ta + Tb}{Ta + Tb + Fa + Fb} \tag{1}$$

Where "Ta" represents true positive values, "Tb" represents true negative values, "Fb" represents false negative values and "Fa" represents values that are false positive.

Definition 2: Precision

Precision metric is defined as the fraction of recovered instances of tweets that are applicable to the query. The formula for calculating the Precision is shown in equation 2:

$$Precision = \frac{Ta}{Ta + Fa} \tag{2}$$

Definition 3: Recall

Recall metric is the fraction of instances of tweets relevant to the query, that are successfully recovered.

The formula is shown below:

$$Recall = \frac{Ta}{Ta + Fb} \tag{3}$$

The accuracy of clustered tweet instances obtained with MFTCT (Modified Fastest Threshold Clustering Technique) is 81.1% when applied on Twitter samples whereas EM produced only 64.3% and K Means produced 75% accuracy, FTCT produced approximately 79.8% accuracy. Nevertheless the proposed method surpasses these three existing clustering techniques with an overall accuracy value of 81.1%. Precision and Recall values shown by our proposed method are 0.57 and 0.33 respectively whereas FTCT shows 0.55 and 0.36 values of precision and recall. K Means clustering method shows 0.55 Precision value and 0.31 as the Recall value. EM method shows approximately 0.43 Precision value and 0.33 Recall value. The metric representation is demonstrated in table 1. The graphical representation of the metric values with the corresponding clustering methods is also depicted below in figure 2 and figure 3:

The empirical results clearly show that our proposed method is more efficient in accordance with the Accuracy, Recall and Precision measures compared to the existing Unsupervised clustering methods.

Table 1
The accuracy representation of clustering methods

Method	*Accuracy*	*Precision*	*Recall*
MFTCT	81.1 %	0.57	0.33
FTCT	79.8%	0.55	0.33
K Means	75%	0.55	0.31
EM	64.3%	0.43	0.33

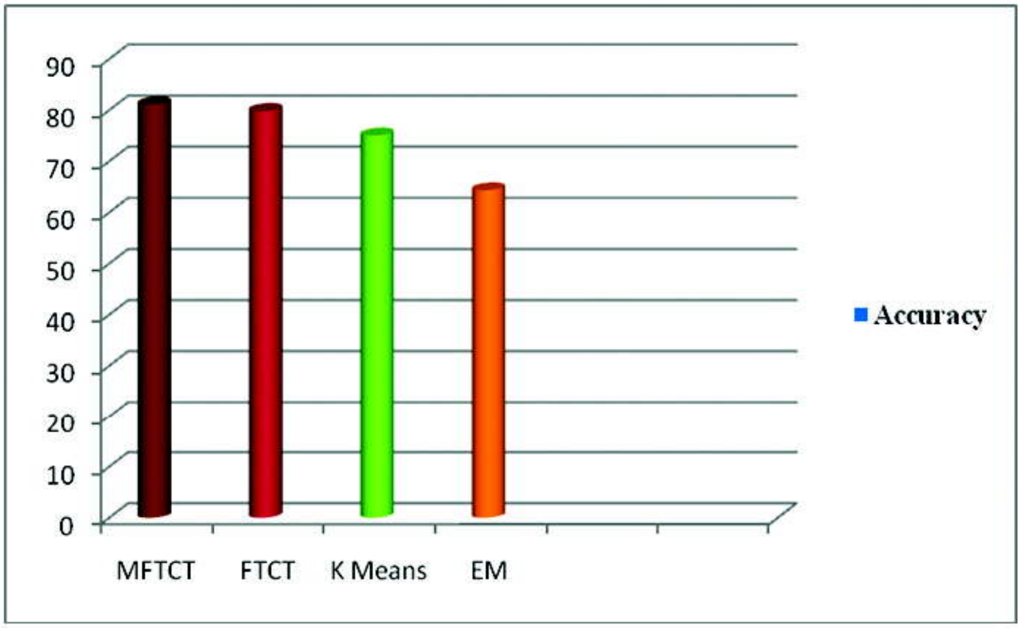

Figure 2: Thegraphical representation of accuracy measures

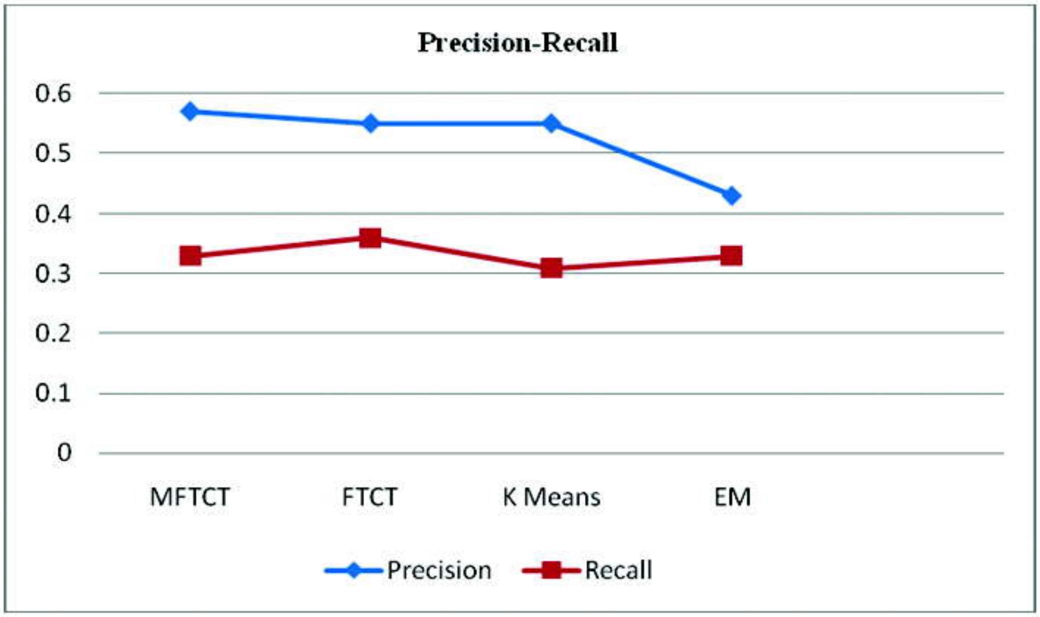

Figure 3: Thegraphical representation of precision & recall measures

5. CONCLUSION

Twitter as a means to distribute critical information has gained much interest as a new social networking site both in terms of users and the amount of engagement. In this paper we have designed and implemented, a novel Unsupervised clustering method and also made a comparative analysis of its effectiveness with other three traditional clustering methods. The Unsupervised clustering techniques provides accurate result with no manual processing or consecutive training. According to the empirical analysis results, the proposed MFTCT method is proven to be suitable in attaining high quality outcomes in terms of evaluation metrics namely Accuracy, Precision and Recall. The proposed method; Modified Fastest Threshold Clustering Technique (MFTCT) gives pretty good results with approximately 81.1% accuracy compared to the existing traditional clustering methods.

REFERENCES

[1] A. Khabia, M. B Chandak, “A Cluster based Approach for Classification of Web”, *International Journal of Advanced Computer research,* pp. 934-938, Dec. 2016.

[2] G. Li F. Liu, “Sentiment analysis base on clustering: a frame work in improving accuracy and recognizin neural opinions,” Springer, pp. 441-452, 2013.

[3] B. Burscher, R. Vliegenthart, H. Claes, “Frames Beyond Words: Applying Cluster and Sentiment Analysis to News Coverage of the Nuclear Power,” *Social Science Computer Review*,SAGE, pp. 1-16, 2015.

[4] L. Zhao, M. Huang, H.Chen, J. Cheng, X.Zhu, “Clustering Aspect related Phrases by leveraging Sentiment Distribution Consistency”, Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1614-1623, Oct.2014.

[5] N. R. Shikalgar, A. M. Dixit, “JIBCA: Jaccard Index based Clustering Algorithm for Mining Online Review”,*International Journal of Computer Applications*, Vol.105, No. 15, Nov. 2014.

[6] L.Zhu, A. Galstyan, “Tripartite Graph clustering for Dynamic Sentiment Analysis on Social Media”, SIGMOID’14, pp. 1531-1542, June. 2014.

[7] S. Part, “Sentiment Classification using Sociolinguistic Clusters”, CEUR, ceur-ws.org, pp. 99-104, 2015.

[8] L. N. Ferreira, L. Zhao, “ A Time series Clustering Technique based onCommunity Detection in Networks”, *Proceedia Computer Science*, Elsevier, pp. 183-190, 2015.

[9] A.Shukla, R. Misra, “ Sentiment Classification and Analysis using Modified K-Means and Naïve Bayes Algorithm”, *International Journal of Advanced Research in CS and Software Engineering*, Vol. 5, Issue 8, Aug. 2015.

[10] L .F .S. Coletta, N. F. F da Silva, E. R. Hruschka, E. R. Hruschka, “ Combining Classification and Clustering for Tweet Sentiment Analysis”, BRACIS (Brazilian Conference on Intelligent Systems) , July. 2016.

[11] N. Pooranam, G. Shyamala, “ A Statistical Method of Knowledge Extraction on Online Stock Forum Using Subspace Clustering with Outlier Detection”, *International Journal of Innovative Research in Science, Engineering and Technology*, Vol.5, Issue 5, May.2016.

[12] M. Unnisa, A. Ameen, S. Raziuddin, “ Opinion Mining on Twitter Data using Unsupervised Learning Technique “, International Journal of Computer Applications, Vol. 148, No.12, pp.12-19, 2016.

[13] A. Sarker, G. Gonzalez, “ Diego Lab 16 at Sem Eval- 2016 Task 4: Sentiment Analysis in Twitter using Centroids, Clusters and Sentiment Lexicons, “ Proceedings of Sem Eval-2016”, pp. 2019-214, 2016.

[14] Lloyd, S, “ Least squares quantization in PCMAC”, *IEEE Transactions on Information Theory*, pp. 129-137, 1982.

[15] N. Li, D. D. Wu, “ text mining and sentiment analysis for online forums hotspot detection and forcast”, Elsevier, pp. 354-368, 2010.