International Journal of Pure and Applied Mathematics
Volume 117 No. 15 2017, 227-239
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

Social Media Data Driven Determination of Student Perceptions

[1]Subhash Bhagavan Kommina and [2]Jinil Persis Devarajan
[1]Hindustan Institute of Technology and Science,
Chennai, India.
subhash@sasi.ac.in
[2]School of Computing Sciences,
Hindustan Institute of Technology and Science,
Chennai, India.
djinil@hindustanuniv.ac.in

Abstract

Social Media is an open platform where every individual can view and share their opinion about any context. Twitters, Facebook, LinkedIn are few to name these. Opinion mining is a process to identify the opinion of the user over a particular context for which Sentiment analysis is used to identify the hidden opinions in the text. Identifying the hidden sentiment from the text is challenging task that requires natural language processing and machine learning techniques. Students share their open feelings about their education in social media sites out of class room. By collecting this kind of information we can identify and investigate the problems of students towards achieving quality higher education. It helps the educational organizations to identify the problems and can improve the stakeholder's satisfaction. However the data collected from the social media is huge and analyzing such data is a challenging task. In this paper we propose a hybrid classifier to classify the text and to identify the students' problems in higher education field.

Key Words: Sentiment analysis, micro blogging, hybrid classifier, social media data.

1. Introduction

Social media networks are very popular in recent years; Now-a-days millions of users share their opinions in social media. Social media can be a reliable source to gather public opinions and factual information in many fields. Twitter is a micro blogging service that allows user to post messages called tweets. These are short messages (140 characters in length), uses emoticons, acronyms and others that express special meanings. Micro blogging websites have become a source of variety of information with millions of messages posted daily on popular websites. Users can post real time messages about their life and share their opinions on variety of topics. Product reviewing will help Companies to find the strengths, weakness and limitations of their products when compared to their competitors.

The Writer's attitude towards a particular topic, or product or service can be computationally identified and then opinions can be categorized (positive, negative or neutral) using Sentiment Analysis. Sentiment Analysis was first introduced in 1990 and it became a major research topic in 2000. In Sentiment Mining there are two kinds of textual information: Facts and Opinions. Facts refer to the objective statements about the product, while Opinion describes the emotion extraction of a product or an issue (Liu, 2010). Sentiment analysis is a sub field of text mining because most of the data is available in text format(Vohra& Teraiya, 2013). Sentiment analysis involves Natural Language Processing, Machine Learning and Artificial Intelligence. Extensive research has been carried out in automatic Sentiment Analysis and the researchers have developed Sentiment classifiers (Bo Pang & Lee, 2004; Karamibekr & Ghorbani, 2012), Opinion extraction (Binali *et al.*, 2009), Sentiment prediction, Text analysis methods and techniques for extracting opinions of customers in the presence of unstructured data.

The rest of the paper is organized as follows: In section 2, the literature review on related work is presented. In the section 3, the flowchart of Sentiment analysis followed by the tools and techniques existing to perform Sentiment analysis is given. In section 4, the proposed methodology along with comparison study of various methods when performed on a sample dataset is presented followed by the authors' recommendations and conclusions are presented in the last section.

2. Literature Review

Related Work

Decision tree is very useful classification and regression technique. Decision Trees are extremely adaptable, straightforward, and simple to investigate and are found to handle characterization and relapse issues(Adinarayana, 2016). They can handle categorical variables and continuous variables and allow us to foresee a persistent quality trees. Decision tree has provision to derive

classifiers from a table of information without much difficulty in pre-processing. Random Decision Tree (RDT) algorithm is more exact than other choice based tree solutions in which the structure of an arbitrary tree is built totally autonomous of the preparation data(Vaidya*et* al., 2014). The RDT calculation can be broken into two stages, preparing and arrangement. The preparation stage comprises of building the trees and populating the hubs along with preparing examples.

Initially tweets are collected from a Social Media, Twitter. Opinion Miner (Liang & Dai, 2013), a process of mining extracts the opinion tweets by employing pre-processing steps. The extracted tweets were then classified for further processing. To improve the overall accuracy rate, the tweets are categorized using text classification. Training data of diverse groups are used to construct classifiers. Significance of words correlated to dissimilar domains are taken out using Short Text Classification (Liang & Dai, 2013). The unigram naïve bayes classifier in conjunction with the pre-labelled training data is used to build the multi-classifier. It needs distinct categories of training data. Significant features can be derived from the entire set of features by relating feature selection algorithms that uses Mutual Information. This ensures that for every class *C* and feature *F*, there exists a score to calculate the contribution of *F* to build a correct conclusion on class *C*.

The Naive Bayes classifier is a prominent system among the essential content arrangement systems with dissimilar applications in email spam management, confidential mail sorting, record categorization, and dialect disclosure and estimation revelation.

Bayes performs well in several troublesome demonstrable difficulties. Naive Bayes classifier is incredibly creative as it is less computationally and it requires a little measure of readiness data Despite the fact that it is much of the time beat by different strategies, for example, helped trees, Max Entropy, Support Vector Machines and so on,. One all around enjoyed approach to execute multi-mark classifier is to change over the multi-name association issue into numerous single-name classification issues(Go *et al.*, 2009). Binary classification presents two categories of classes.

In Single label classification, classes are mutually exclusive and each tweets fall into one of these classes.(Wilson *et al.*, 2009; Khairnar & Kinikar, 2013). Most active studies found on tweet planning are either paired order on significant and insignificant matter, or multi-class grouping on non definite classes, for example, news, occasions, sentiments, arrangements, and confidential messages.

Notion investigation is a different extremely widespread three-class arrangement on positive, negative, or neutral deductions(B. Pang & Lee, 2004). Client assessments on products can be mined through the survey or online posts. Slant examination helps in this case.(Sagar*et al.*, 1996).Abundant systems are

formed to quarry emotions from texts. But, just studying the slant of standby posted tweets does not give much remarkable learning on significant mediations and supervisions for understudies.

In this process, Natural language processing is used to identify and retrieve certain information from the text or tweet. This is done by identifying subjectivity from the text posted by users(Blair-Goldensohn*et al.*, 2008). For example, to identify the success of a product it is very important to identify the product features and to analyze which feature the user mostly liked or disliked, and this process is termed as feature extraction from text. Sentiment analysis techniques are also used to identify the polarities of the text, either it can be positive, negative or neutral. By doing these kinds of analysis we can identify which feature of the product is mostly liked by the users and which is not(Cámara*et al.*, 2013).By analysing the polarities, the good features and the bad features can be derived out depending on the weight of the word. Then sentiment of the user over a particular product can be obtained by capturing his/her emotion emotions such as joy, frustrated and neutral.

In this study, the sentiments of students who are using social media to share their feelings about the quality of education provided by a college is considered. Students' emotions are very important to identify the problems related to their learning process. By analyzing the tweets posted by students it is easy to identify the corresponding problems faced in their learning activity. Qualitative analysis of this data on social media will help to identify the problems associated with their education system. The feedback thus obtained can be given back to the education system and improvement can be made.

There are two main approaches for Sentiment Analysis, namely Bag of Words approach and Feature Based Sentiment (Tan *et al.*, 2014). In Bag of Words approach, the syntax and semantics are given which could not be used for opinion mining. Feature Based Sentiment approach is used for analyzing the sentiment of products, services etc. Identifying the sentiments from user generated data on micro blogging websites has been used in the literature(Khairnar& Kinikar, 2013). If the text is written in digital format, from the words sentiments can be detected(Blair-Goldensohn*et al.*, 2008).

In this study, different kinds of classification algorithms commonly used in literature to perform sentiment analysis are discussed and compared. Also sentiment analysis on textual tweets from twitter data about education is carried out using SVM classifier and the performance is compared with the existing algorithms.

3. Sentiment Analysis

The stepwise procedure in the process of sentiment analysis is given in Fig. 1.

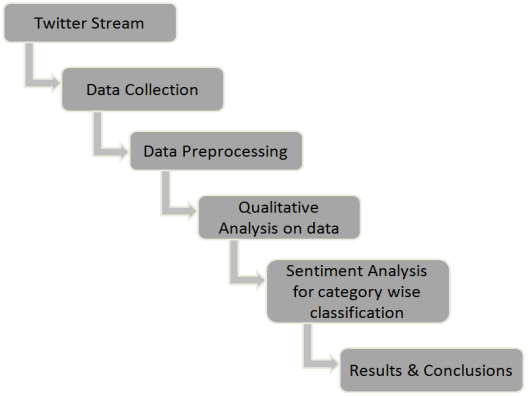

Figure 1: Architecture Workflow

Steps in Sentiment analysis are:

1. Data Source: The data is downloaded from twitter with the support of Twitter API. Twitter provides two kinds of API. They are search API and Stream API. Data can also be downloaded from UCI, Kdnuggets and SNAP.
2. Text Extraction – This step includes separating words from content that impact the result of the outcome. The data will be stored in a csv or excel files.
3. Text Refinement – This step includes refining the content into important expressions, words and so forth. In this the raw data is collected and pre-processed by removing the Hash tags, re tweets, stop words, symbols, and uniform casing and slang words.
4. Text Classification – This step incorporates grouping of content into its class (positive/negative). Here the corpus is constructed that consists of all the possible words in the data set. It treats every tweet as a document and classifies each document into specific class.
5. Score Aggregation – This step gathers aggregate scores from classifier and further summed up to make the aggregate to produce score *(Ja*mes, *G.*, *W*itten, D., Hastie, T., Tibshirani, 2013).

Sentiment Analysis Tools and Techniques

The mostly used techniques to perform sentiment analysis are Naive Bayes classification (NB), Support Vector Machines (SVM) and Maximum Entropy (ME). Naive Bayes performance is good and accurate when it is implemented on reviews and blogs. But when it comes to Lexical based methods the results varies and performance is not accurate and bad. Literature reveals that the performance of SVM is good and high when compared to NB using Lexical based approach(Wilson *et al.*, 2009). During analysis, SVM performs well when compared to NB and ME. So, SVM classifier is believed to be a better classifier when compared to remaining Lexical based approaches though there are some differences in the overall performance.

1. Naive Bayes Approach

Naïve Bayes is a simple model which is used for text categorization. The Naive Bayes classifier depends on Bayes hypothesis

Class c* is assigned to tweet d, where

$$c *= argmaccPNB(c|d)$$

$$PNB\ (c|d):= \frac{P(c)\sum_{i=1}^{m} P(f|c)^{n_i(d))}}{P(d)} * P(d|c)$$

Where P(c) and P(d) are prior probabilities of a class. Naïve bayes classifier can find the probability of a word belonging to a particular class or not ("Machine Learning with Naive bayes classifier," n.d.).

$$\underline{P\ (Xi|c)\ =\ \ \ \ Count\ of\ Xi\ in\ document\ of\ class\ c}$$
$$Total\ no\ of\ words\ in\ document\ of\ class\ c.$$

2. Maximum Entropy

It is a probability distribution estimation technique(Cámara*et al.*, 2013) and a featured based model and it is one of the machine learning algorithms. It depends up on probability approach like naïve bayes approach. It is used when most of the data is unknown but the data is uniform. It is given as.

$$P(c|d)\ =\ 1\ Z(d)\ exp(X\ i\ \lambda ifi(d,c))$$

3. Support Vector Machine

The support vector machine was the most complex algorithm and it is a common approach for text classification (Pak & Paroubek, 2010). It is popular because of its high accuracy. The support vector machine is classed as a non-probabilistic binary linear classifier. It works by plotting data in multidimensional space. It then tries to separate the classes with a hyper plane, hp. If the classes are not separated directly then add new dimension in the multidimensional space. The calculation includes another measurement trying to further separate the classes. The process is continued until it has the capacity to isolate the training data into its two separate classes using a hyper plane.

If a hyper plane hp is drawn, and the tweet, t, corresponds to class $C_i \in \{1,-1\}$ in which the tweet has to be classified.

$$y(x)\ =\ sign[\ X\ \sum_{k=1} N\ \alpha k\ yk\ \psi(x,xk\ )\ +\ b]$$

Where $x_i = \{x_{i1}, x_{i2}, \ldots . x_{in}\}$ is a input vector where $y_i$is a output class when the text is linearly separable. In this study, student's emotions are classified whether positive or negative. The proposed algorithm classifies student's problems into different categories and preventive measures can be taken to avoid such problems. The algorithm also incorporates Fuzzy Logic techniques for handling the ambiguous and imprecision words of a tweet.

4. Proposed Hybrid Classification Method

In this method, the tweets are collected as tokens and cleaned using text normalization process. Fuzzy logic is applied to remove if any ambiguity between tweets. Further SVM is used to classify the tweets into corresponding categories.

Algorithm

Input: A dataset T={t_1,t_2,t_3,. . . . .,t_n} of n tweets which are downloaded from # engineering problems, #engineering perks, #college problems, #lady engineer.

Output: Classification of tweets into categories like C_1,C_2,C_3. . . .C_y.

Hybrid Classifier (T)

Begin:

1. Collect the tweets into a data frame.
2. Apply tokenization process where each tweet is divided into individual words.
3. Apply word normalization process by converting raw data to cleaned data.
4. Remove the notations, symbols, URLs, stop words, emoticons.
5. Create a corpus using lexicon based method for the classes generated.
6. Identify the most frequent words and create a word vector.
7. Apply Fuzzy logic to the tweet to remove the ambiguity in the tweets
8. Apply SVM to classify the unambiguous tweets for their corresponding class categories
9. Check the scores for each tweet for a specific class category
10. Display results.

End.

In this approach, initially tweets are collected from # engineering students, #engineering problems and #lady engineer as in step 1. The data collected is pre processed and then converted into uniform case. Following this, some twitter notations like hash tags, retweets, symbols, numbers and stop words are removed from step 2-4 is carried out. Then a corpus is created to find the sentiments of the tweets using lexicon based approach as given in step 5. Additionally fuzzy techniques are applied to remove the ambiguity in tweets to make the tweet fall under a specific category as specified in step 6. Further, SVM is employed to categorize the tweets to a specific class basing on the scores of the polarities of tweets as given in step 7-8. Ultimately, the results generated help to identify the most frequent words and the associated problems. Succeeding students during their learning process will get benefited in improving the quality of education (Kinchin, 2017)

Comparison Study

In this study, a sample data set has been collected with the tweets from twitter by specifying the hash tags like #engineering students, #engineering problems

and #lady engineer and sentiment analysis is conducted. Different algorithms re applied on this data set. The performance of SVM algorithm is found better than the other machine learning algorithms in the experimental study. The accuracy of the algorithm is calculated using the performance measures. To measure the accuracy of the model and for predicting the sentiment in the tweet, the following parameters: True-Positive (TP), True- Negative (TN), False – Positive (FP) and False-Negative (FN) are taken into consideration for constructing the confusion matrix. In sentiment analysis, the performance measures to be considered are Precision, F-measure, Recall are used for evaluating the model in predicting the polarity of the tweet (Adinarayana, 2016).

The Precision (P) Measure is given as

$$Precision\ (P)\ = TP/\ (TP + FP)$$

The Recall (R) Measure is given as

$$Recall(R) = TP/(TP + FN)$$

The F-score (F) Measure is given as

$$F - score\ (F)\ =\ (2 * P * R)/\ (P + R)$$

Table 1: Precision of Algorithms in Cross Validation of the Tweets

Algorithm	Precision of algorithm (in %)	Recall of algorithm	F-score of algorithm
Boosting	63.6	66.6	64.6
Random Forest	46.8	62.5	51.6
TREE	63.6	66	64.6
SLDA	59.6	58	61
Hybrid Classification	72.4	68.3	68.6
BAGGING	63	66.6	64.6

Figure 2 shows the precision, recall and f-score of the algorithms in identifying the problems effectively. It is found that the SVM based hybrid classifier is able to identify the problems more accurately when compared to other methods.

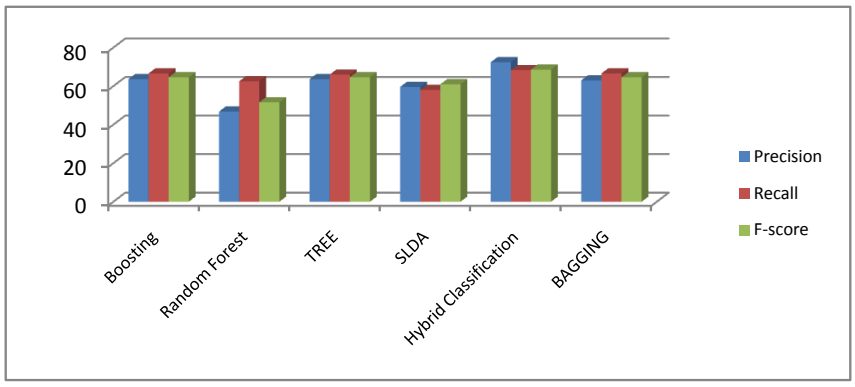

Figure 2: Precision, Recall and F-Score of Algorithms in Cross Validation of Tweets

5. Results

The analysis of these tweets reveals that, most of the engineering students are suffering through the education system. 80% of the tweets are negative tweets and 20% positive. The histogram of most frequent words in the tweets is given in Fig. 3. Inference from this analysis is that the students suffer from distress, work pressure and assignments.

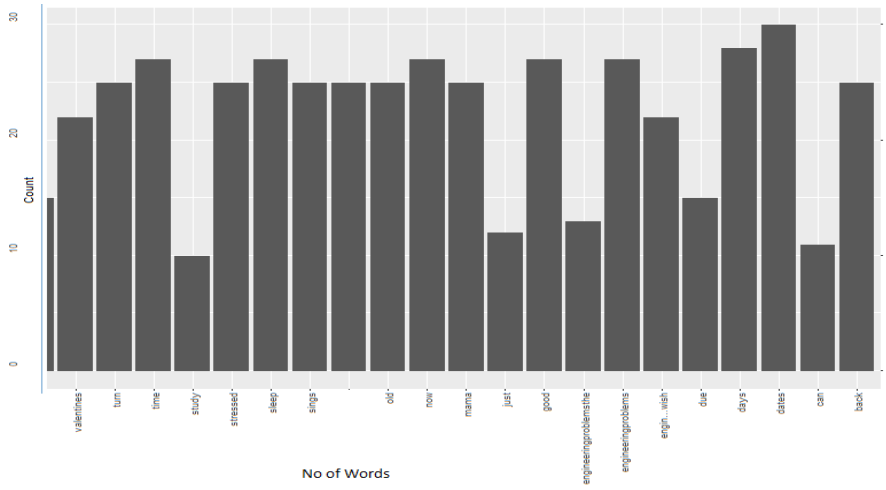

Figure 3: Histogram for Most Frequent Words

From these words we can identify various emotions of the students which is given in Fig. 4 and the tweets are classified accordingly.

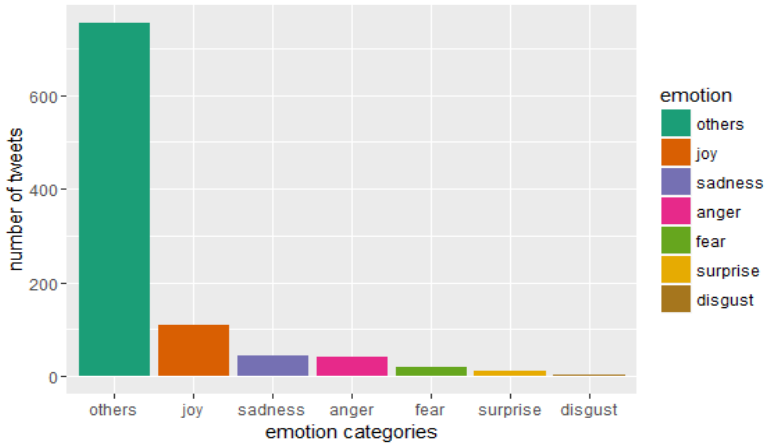

Figure 4: Emotions of Engineering Students

From the analysis, it is clear that SVM performance is effective compared to NB since it is able to bring out more number of classes from the given input as shown in Fig. 5.The classification accuracy of this hybrid classifier is more than the Naïve based classifier.

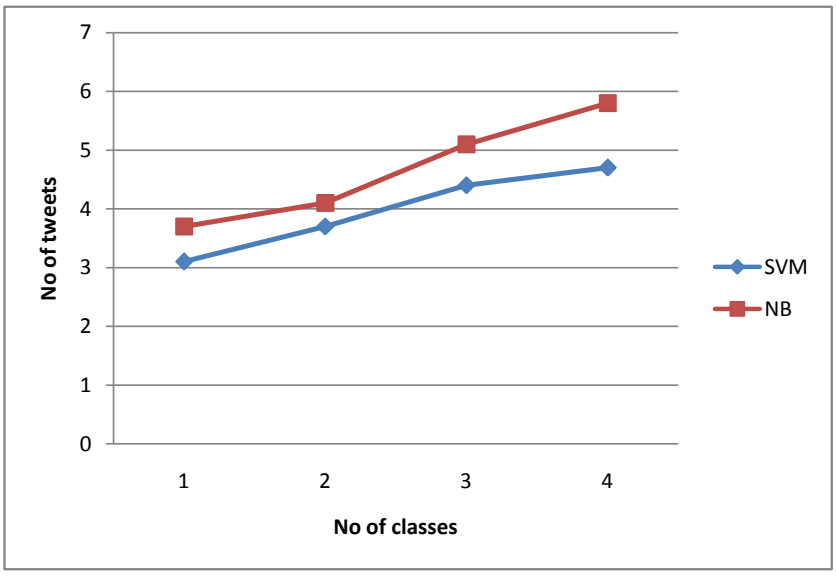

Figure 5: Hybrid Classifier Vs. NB w.r.t. Number of Classes

6. Conclusion and Future Work

Through the evaluation of recent social network analysis work, especially on public opinion mining up-to-date, this paper has identified and emphasized the importance of sentiment analysis. The proposed methodology for social media mining, namely, Hybrid Classifier, has succeeded in identifying the students' problems with experimental results to help the educational organisation to take relevant decisions to mitigate the problems. A fair amount of attention is required to perform large scale analysis and to improve the accuracy of the classifier. Our future work is to analyze maximum number of tweets and identify the emotions in other fields such as hotel industry, health care services, product industry and so on to discover connected problems that needs to be tracked for quality improvement.

References

[1] Adinarayana S., Ilavarasan, E., Classification techniques for sentiment discovery-A review, IEEE International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES) (2016), 396-400.

[2] Binali H., Potdar V., Wu C., A state of the art opinion mining and its application domains, IEEE International Conference on Industrial Technology (2009), 1-6.

[3] Blair-Goldensohn S., Hannan K., McDonald R., Neylon T., Reis G.A., Reynar J., Building a Sentiment Summarizer for Local Service Reviews, WWW Workshop on NLP in the Information Explosion Era (2008), 339–348.

[4] Cámara E.M., Valdivia M.T.M., López L.A.U., Ráez A.R.M., Sentiment analysis in Twitter Sentiment analysis in Twitter, Natural Language Engineering 20 (1) (2013), 1–28.

[5] Kinchin I.M., Correia P.R., Pedagogic frailty and concept mapping, Knowledge Management & E-Learning An International Journal (KM&EL) 9(3) (2017), 254-260.

[6] Go A., Bhayani R., Huang L., Twitter Sentiment Classification using Distant Supervision, Processing 150 (12) (2009), 1–6.

[7] James G., Witten D., Hastie T., Tibshirani R., An Introduction to Statistical Learning with Applications in R (2013).

[8] Kinchin I.M., Pedagogic frailty: A concept analysis. Knowledge Management & E-Learning: An International Journal (KM&EL) 9(3) (2017), 295-310.

[9] Karamibekr M., Ghorbani A.A., Sentiment Analysis of Social Issues, International Conference on Social Informatics, (Social Informatics) (2012), 215–221.

[10] Khairnar J., Kinikar M., Machine Learning Algorithms for Opinion Mining and Sentiment Classification, International Journal of Scientific and Research Publications 3(6) (2013), 1–6.

[11] Liang P.W., Dai B.R., Opinion Mining on Social Media Data, 14th International Conference on Mobile Data Management (2013), 91–96.

[12] Liu B., Sentiment Analysis and Subjectivity, Handbook of Natural Language Processing (1) (2010), 1–38.

[13] Jason B., Machine Learning with Naive bayes classifier. (n.d.), (2016).

[14] Pak A., Paroubek P., Twitter as a Corpus for Sentiment Analysis and Opinion Mining, In Proceedings of the Seventh Conference on International Language Resources and Evaluation (2010), 1320–1326.

[15] Pang B., Lee L., Opinion Mining and Sentiment Analysis, Found. Trends Inf. Retr. 2 (2004).

[16] Pang B., Lee L., A Sentimental Education: Sentiment Analysis using Subjectivity Summation based on Minimum Cuts, ACL ’04 Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics 271 (2004).

[17] Sagar V.K., Chong S.W., Leedham C.G., Solihin Y., Slant manipulation and character segmentation for forensic document examination, In IEEE TENCON, Digital Signal Processing Applications, (1996).

[18] Tan S., Li Y., Sun H., Guan Z., Yan X., Bu J., Chen C., He X., Interpreting the public sentiment variations on Twitter, IEEE

Transactions on Knowledge and Data Engineering 26(5) (2014), 1158–1170.

[19] Vaidya J., Shafiq B., Fan W., Mehmood D., Lorenzi D., A Random Decision Tree Framework for Privacy-Preserving Data Mining, IEEE Transactions on Dependable and Secure Computing 11(5) (2014), 399–411.

[20] Vohra S.M., Teraiya J.B., A comparative Study of Sentiment Analysis Techniques, Journal JIKRCE 2(2) (2013), 313–317.

[21] Wilson T.A., Wiebe J., Hoffmann P., Recognizing Contextual Polarity: an exploration of features for phrase-level sentiment analysis, Computational Linguistics 35(3) (2009), 399–433.

[22] RAJESH, M. "A SYSTEMATIC REVIEW OF CLOUD SECURITY CHALLENGES IN HIGHER EDUCATION." The Online Journal of Distance Education and e- Learning 5.4 (2017): 1.

[23] Rajesh, M., and J. M. Gnanasekar. "Protected Routing in Wireless Sensor Networks: A study on Aimed at Circulation." Computer Engineering and Intelligent Systems 6.8: 24-26.

[24] Rajesh, M., and J. M. Gnanasekar. "Congestion control in heterogeneous WANET using FRCC." Journal of Chemical and Pharmaceutical Sciences ISSN 974 (2015): 2115.

[25] Rajesh, M., and J. M. Gnanasekar. "Hop-by-hop Channel-Alert Routing to Congestion Control in Wireless Sensor Networks." Control Theory and Informatics 5.4 (2015): 1-11.

[26] Rajesh, M., and J. M. Gnanasekar. "Multiple-Client Information Administration via Forceful Database Prototype Design (FDPD)." IJRESTS 1.1 (2015): 1-6.

[27] Rajesh, M. "Control Plan transmit to Congestion Control for AdHoc Networks." Universal Journal of Management & Information Technology (UJMIT) 1 (2016): 8-11.

[28] Rajesh, M., and J. M. Gnanasekar. "Consistently neighbor detection for MANET." Communication and Electronics Systems (ICCES), International Conference on. IEEE, 2016.