International Journal of Latest Trends in Engineering and Technology
Vol.(7)Issue(3), pp. 153-161
DOI: http://dx.doi.org/10.21172/1.73.521
e-ISSN:2278-621X

E-OPTIMA(Enhanced-Opinionated Tweet Implied Mining and Analysis) An Innovative Tool to Automate Information Credibility Analysis

Ram Chatterjee[1] and Samiksha Agarwal[2]

Abstract- The predilection and efficacy of the web has procreated significant and magnificent data in the arena of online social networks, impacting criticality of Social Data Analytics profusely. This inclination of expressing opinions, attributed by its diversity, produces voluminous online information that is questionable in terms of its veracity. As different individuals attempt to spread gossipy tidbits, generally only for the sake of entertainment and/or spreading rumors, it is required that there exist strategies utilizing which individuals can, without much hassle, check data believability of information posted on online networks. The recent research endeavors done have been essentially centered around following and displaying opinions of individuals. This paper proposes approach for mechanizing the procedure of congregating opinions posted on Twitter, probing the validity of these beliefs, and assessing its credibility, by implementing the strategies for Sentiment Analysis and Information Credibility Analysis. This implicates assimilation of Latent Dirichlet Allocation (LDA) calculation for grouping of subjects inside the tweets and semi-directed Support Vector Machine (SVM) for sentiments' investigation. As a last point, dominant part choice, by looking at the quantity of Contrastive Opinions around a point is performed, for Credibility Analysis. To address the purpose, an innovative tool, to automate information credibility analysis has been developed, named as E-OPTIMA (Enhanced- OPinionated Tweet Implied Mining and Analysis) ver. 2.0.0. which is an enhanced version of the tool OPTIMA ver. 1.0.0. that caters to automation of sentiment analysis and the results have been presented in a graphical form and tabular manner for ease of understanding.

Keywords –Analysis, Contrastive Opinions, Latent Dirichlet Allocation, Sentiment Analysis, Support Vector Machine, Tweets.

I. THE PREAMBLE

"Twitter", being one of the most coveted medium of online social networking, experiences an unstable and voluminous growth of messages and tweets, being posted freely and frequently. This connotes the need to break down the information on Twitter for authenticating the believability of this information. Apropos to this scrutiny of information posted on online networking sites, the tactics of Sentiment Analysis connotes to inspection of feelings or slants of individuals over a specific context, which is being catered by the field of Social Data Analytics [1]. It can likewise be considered as implication of Natural Language Processing (NLP), which tracks the state of mind of an entity around a specific item or topic [2], and is coined as "Feeling mining" and "Sentiment examination". However, tweets posted may be untrue, useless, unimportant and/or uproarious, that spreads foundation points which are totally unimportant for thought. This necessitates the integration of FB-LDA approach to isolate important points from immaterial ones [3].

[1] Department of Computer Science and Technology Manav Rachna University (Formerly MRCE) Faridabad, India
[2] Manav Rachna University (Formerly MRCE) Faridabad, India

Information Credibility incorporates techniques for discovering and discriminating information with high believability from the morass of information that is accessible, and hence, is a distinguished domain amongst the most vital undertakings in the realm of social information investigation. This can be attributed to the voracious accumulation of new information over and above the existing volume and varied sources of information, creating a mess of ever growing and over growing information, making the task of information credibility analysis, a complicated herculean task. Although varied research have been completed for identifying the false gossipy tidbits on the web [5] [6], however, they are particular to elements of target website pages, for example, the quantity of clients or the structure of site page and so on and so forth. **It is here where the tool "E-OPTIMA" finds its role and significance by serving to fulfill the need of making a better and befitting decision by a customer/user in the context of acquiring the product, process and/or service, based on its credibility analysis.** The legitimacy of credibility in this paper has been figured by ascertaining the proportion of same conclusions concerning each of the feelings that preserves about the point. For finding the topic referred in the tweet, the improved variant of LDA algorithm has been utilized, which is implemented after pre-processing the tweets. Semi- supervised machine learning algorithm, "Support Vector Machine", has been integrated for sentiment grouping.

II. THEORETICAL UNDERPINNINGS OF TWEET RETRIEVAL AND ANALYSIS

In particular, this paper emphasizes on the automation of Credibility Analysis on "tweets" and it is for the readers interest and knowledge we specify the twitter relevant fundamentals. Twitter permits just 140 character announcements (prominently known as "tweets") and has different extraordinary properties like hashtags, targets, emoticons, and so on. In addition, the twitter information is accessible publicly for examination; and there are APIs which can pull information from twitter site. There are plenty of procedures accessible for the tweet recovery which can be extensively sorted as "programming" and "non-programming" systems. The programming methods incorporate 'R' programming language which is a factual language for tweet recovery and its examination. 'R' language can be effectively interfaced with java, C, C++ utilizing the different inbuilt packages of 'R'. Python is another programming language utilized for comparable undertaking. Further, non-programming methods incorporate different online applications named as "Streamcrab", "Topsy", "Notion 140", "Estimation viz", "Trackur" and the likes. These are really time saver applications focusing on the end goal to get what clients need. Clients just need to provide a watchword on which the application focuses for tweets, and in a moment, several tweets are shown on the client's screen. Be that as it may, the appropriateness of the simplicity and adaptability of non-programming strategies, is repudiated by the troubles confronted while sending out tweets into an external document and making them accessible for pre-processing. Hence, for the ease of usability and accessibility, we have chosen 'R' language for recovering and probing the tweets. 'R' being an open source software and can be easily downloaded from the internet [7].

III. INTRICACIES OF THE MACHINE LEARNING ALGORITHMS APPLIED

For combination of the hidden ideas, it is fundamental to say that the point of Machine Learning is to build up a calculation in order to upgrade the execution of the framework utilizing case information or past experience. The Machine Learning gives an answer for the grouping issue that includes two stages:

1) Learning the model from a corpus for preparing information.

2) Classifying the unknown information in light of the prepared model.

The term "artificial intelligence" is likely to be applied when a machine uses cutting-edge techniques to competently perform or mimic "cognitive" functions that we intuitively associate with human minds, such as "learning" and "problem solving". **This paper primarily focuses on automation of two such machine learning algorithms viz semi-supervised SVM and LDA.** The support vector machine is supervised learning model with related learning calculations that dissect information utilized for grouping and relapse examination. There are intuitive ways to solve multiclass with SVM as stated in figure 1 [8].

$$\min_{\mathbf{w}} \frac{1}{2}\sum_{i=1}^{K} w_i^T w_i + \frac{C}{n}\sum_{i=1}^{n} \max_{y \neq y_i} \{0, 1 - w_{y_i}^T x_i + w_y^T x_i\}, \quad (1)$$

Figure 1. Multiclass SVM Formula

where
$\mathbf{w}$ in (1) is a matrix with column w_i
$i \in 1, \ldots, K$.
$w_y^T x_i$ is the confidence score of tweet t_i
belonging to class y.

In general, the performance of sentiment classification is evaluated by using the following indexes. They are Precision, Recall and F1-score. The common way for computing these indexes is based on the confusion matrix [9] as shown below:

#	Predicted positives	Predicted negatives
Actual positive instances	Number of True Positive instances (TP)	Number of False Negative instances (FN)
Actual negative instances	Number of False Positive instances (FP)	Number of True Negative instances (TN)

Figure 2. Confuion Matrix

Precision is the portion of true positive predicted instances against all positive predicted instances. Recall is the portion of true positive predicted instances against all actual positive instances. F1 is a harmonic average of precision and recall. To consolidate the concept discussed so far, we present below these indexes that can be defined by the following equations:

- $\text{Precision} = \frac{TP}{TP + FP}$
- $\text{Recall} = \frac{TP}{TP + FN}$
- $\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

In the purview of NLP, LDA is a generative statistical model that explicates sets of observations apropos to unobserved groups, that elucidates and draws attention as to why some parts of information are analogous. Congenial to this, emphasizing on topic classification (that corroborates analysis of sentiments), enhanced LDA algorithm viz. FB-LDA, has been used, which classifies the tweets as background set and foreground set. Readers may note that background set consists of the set of tweets that refer to the topics which were appearing before the variation period whereas foreground set consists of the set of tweets that entails topics appearing within the variation period, where "variation period" is the time where there is significant increase in either positive or negative tweets about a topic.

Further to this, the technique proceeds with finding out the topics which satisfies these two conditions: a) the topics should be present in foreground set; b) they must be outside the background set. This puts word distribution into use. The process culminates with the implication of the Multinomial distribution function that assigns topic to the tweet, as stated in figure 3.

$$f(x_1, \ldots, x_k; n, p_1, \ldots, p_k) = \Pr(X_1 = x_1 \text{ and } \ldots \text{ and } X_k = x_k) = \begin{cases} \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}, & \text{when } \sum_{i=1}^{k} x_i = n \\ 0 & \text{otherwise,} \end{cases}$$

Figure 3. Multinomial distribution

Where:
n is the number of tweets
k is the number of topics
pi is the probability that the tweet belongs to topic i

IV. SYNOPSIS OF THE TOOL- E-OPTIMA VERSION 2.0.0

A. *Prologue on the Tool*

For the reader to decipher the working of the tool analytically, we commence with the brief introduction on the tool. E-OPTIMA as stated, has been developed in "R" programming language,

and in fact has been portrayed as a web application that consolidates the various phases of Information Credibility Analysis viz. Tweet extraction, Text preprocessing, application of Machine Learning algorithms for Sentiment Analysis, Topic Classification and finally computing the Degree of Credibility.

B. Tool Portrayal

To reveal the specialized intricacies of E-OPTIMA, a firm and precisely formal clarification requires it worth specifying that few "R" packages have been utilized to enquire the Twitter search API to ease the goal of gathering tweets into R for Opinion Mining, by promoting a secured association between the R console and the twitter.

The tool assimilates an in-built function to extract tweets based on the search keyword, given by the user. Additionally, the tool consists of the option to decide on the upper bound of the tweets for extraction. Also, the tool is capable of showing the tweets in tabular format. The tool further generates “word cloud” connoting to the most frequent terms in the “corpus” formed from the extracted tweets. To clarify further, the initial tool interface has been shown in figure 4(a) in labeled form for readers’ clear comprehension.

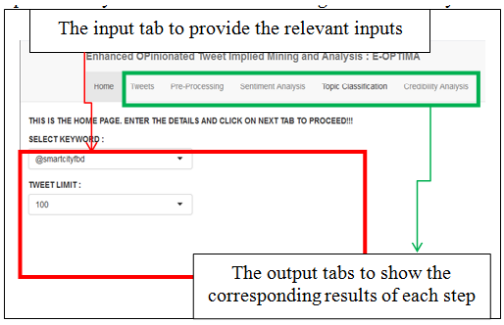
As depicted in the figure, the tool implicates two initial inputs and five output tabs, providing the provision to the user to switch among the tabs in order to view the multifarious outputs produced by the tool in context of the given “select keyword”.

C. Illustrating the tool input and output

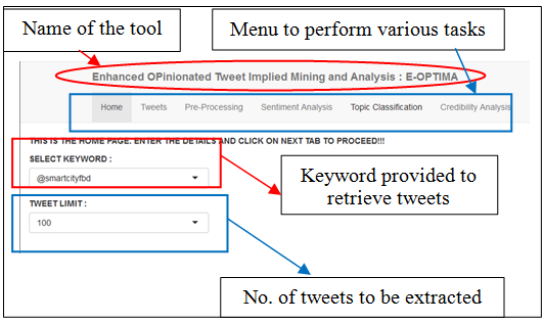
To expound further on the tool’s interface, besides the “select keyword” being the first input denoting the search keyword (which has been set as “@smartcityfbd”), the second input parameter to the tool is the “Tweet Limit” that fixes the upper bound on the number of tweets to be extracted from the twitter site. The tool facilitates leveraging this “Tweet Limit” up to 2500 tweet extraction as the upper bound limit, which may be selected from the associated drop down list, as illustrated in figure 4(b). It is for the readers’ ease of understanding, we state further that the drop down list associated with the “select keyword” input parameter, facilitates the end user to re-select any past searched keyword, on which the end user may like to re-extract relevant tweets at some point of time in the future. This feature fulfills the perspective of re-building the corpus with enhanced number, and quality of tweets, congenial to the updates posted on twitter frequently, randomly and dynamically, in context of any specific topic of interest to the end user. This fulfills the purpose of tool development by making it “generic” in terms of its searching ability, in the context of diversity in content, context and number of tweets, on the twitter site.

Demarcating the outputs from the inputs, we have been depicting the outputs from figure 4(c) to figure 6(b) in self-explanatory manner which are being corroborated with the following detailed discussion. Figure 6 portrays the first output which connotes to the list of retrieved tweets which are the “raw tweets” comprised of stop words, URL’s, punctuation marks ,emoticons and hash tags.

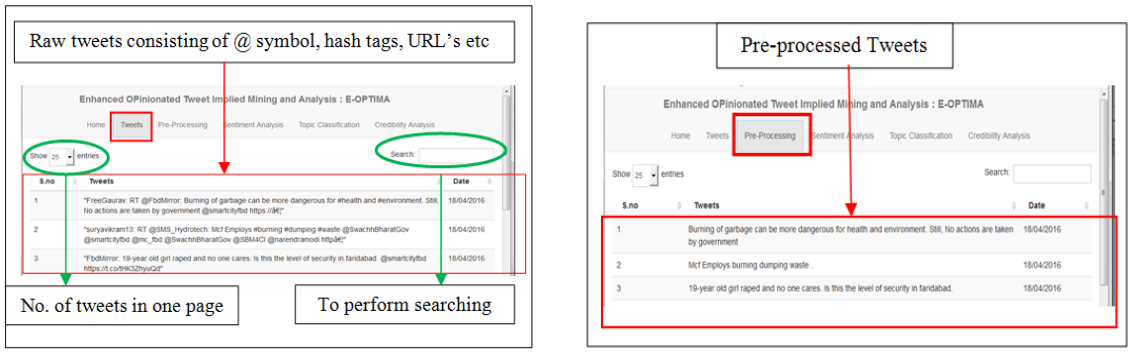
Being in crude shape, these tweets are subjected to content pre-processing [10] where the after effect of content handling are depicted in figure 4(d). Clicking on “pre-processing” tab serves the purpose mentioned above.



(a)



(b)



(c) (d)

Figure 4. (a) Tool Interface (b) The Input Tab (c) The Tweet Tab (d) The Pre-Procesing Tab

Proceeding the discussion further, the next output tab is the tab that performs the function of Sentiment Analysis or Opinion Mining and consists of a sub-menu that shows the results in various forms. The first tab under "Sentiment Analysis" Tab is "Senti-Plot I" which focuses on plotting the graph based on the emotions captured in the tweets as depicted in figure 5(a). Readers are advised to comprehend "unknown" legend depicting the tweets that does not contain any emoticons and the rest depict the emotions such as "joy" "surprise" and "anger". Also, it is advised to comprehend score "1" as positive tweet, score "-1" as negative tweet and score "0" as neutral tweet, where positive, negative and neutral terms refer to the underlying sentiment, inherent in the retrieved tweets. Technically, the "score" is calculated as:

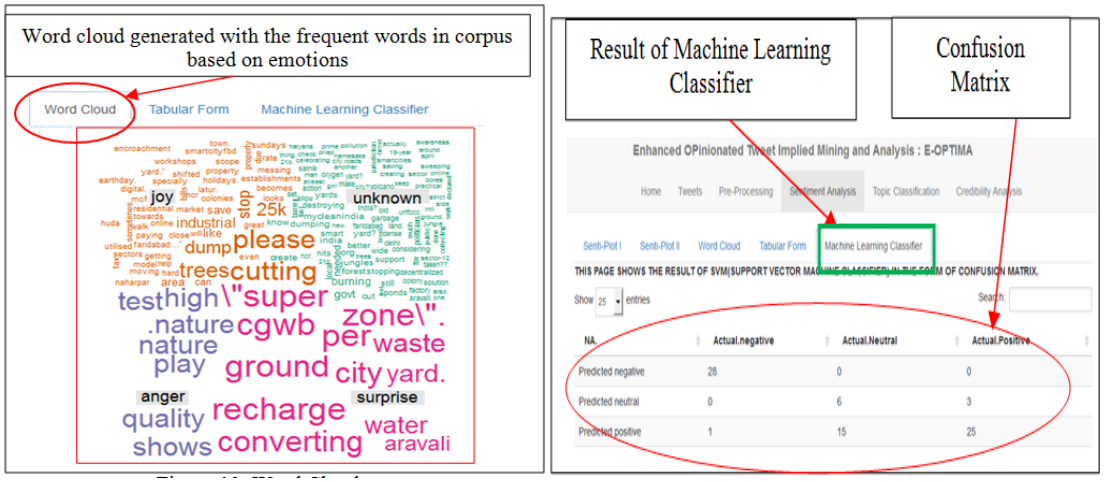
Score= (Sum of Positive words) – (Sum of Negative words)

Figure 5(b), depicts the second tab under "Sentiment Analysis" Tab as "Senti-Plot II" which focus on plotting the graph based on the polarity of tweets.

The results of this classification are exported into a word cloud, as illustrated in figure 5(c), which forms the next output. Word Clouds are graphical depictions of word recurrence that give better significance than words that seem all the more redundant in the source content. The more a word is observed to be regular in the reports, the greater is its appearance in the visual structure of the word cloud. Representation of this kind authenticates evaluators with investigative literary examination.

To establish the accuracy of the retrieved results, machine learning classifier semi-supervised Support Vector Machine (SVM) has been implicated in the tool E-OPTIMA. To achieve the purpose, 30% of the tweets are used as training data and 70% are used as testing data. The classifier generates a "confusion matrix" which serves as the basis for calculating several performance parameters. Figure 5(d) shows the confusion matrix.

(a) (b)



(c) (d)

Figure 5. (a) Plot distribution of emotions (b) Plot distribution of polarity (c) Word Cloud (d) Confusion Matrix

After the completion of Sentiment Analysis, the next step is to apply Topic Classification whose results have been shown in figure 6(a). The "Topic Classification" tab is divided in 2 more tabs namely "Topic Classified" and "Topics Assigned to Tweets". The result of LDA algorithm combined with FB-LDA gives the List of topics encountered in tweets and associates each tweet with a topic. The last stride is the calculation of information credibility score which is computed in the light of the following equation:

$$\text{Credibility} = \frac{\text{number of tweets referring to same opnions about the topic as given tweet}}{\text{total number of tweets having opinion on the same topic as that of a given tweet.}}$$

Finally, the results are shown in both tabular as well as graphical form which are depicted in figure 6(b).

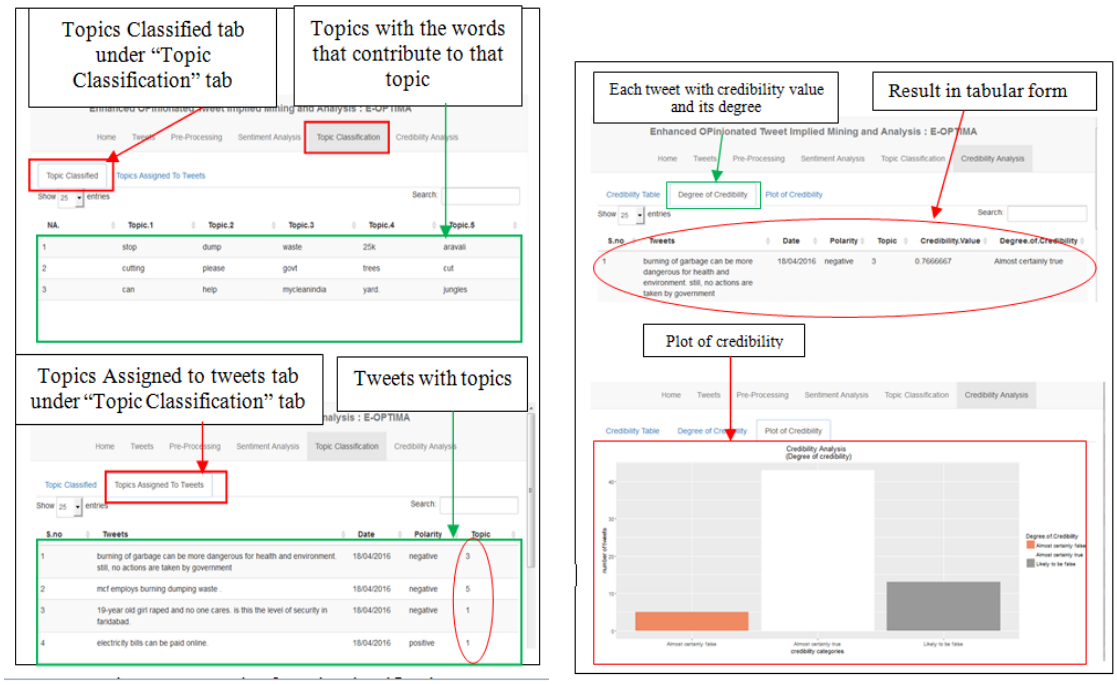
D. Detailing Deployment of the Tool

To consolidate the idea of the yields obtained so far, it is for the readers' appreciation, we explain "Credibility" further.The degree of credibility is assigned based on the following table.

Table -1 Deriving Credibility Of Information [4]

Calculated Value	**Degree of Credibility**
Greater than 0.7	Almost certainly true
0.3 - 0.7	Likely to be false
Less than 0.3	Almost certainly false

Considering this table as the basis, the degree of credibility has been assigned to each of the tweets.



(a) (b)

Figure 6. (a) Result of Topic Classification (b) Result of Credibility Analysis

E. Outlining the Tool Output

Technically speaking, the critical aspect lies in automating credibility analysis procedure by mechanizing the procedure of discovering valid data out of the entire parcel of data accessible on the online social network "twitter". And, this cannot be achieved by just applying the equation. Improvements must be done amid the procedure in order to pick up trust in the outcome. This improvement, attributed to SVM and LDA have been automated in the tool E-OPTIMA.

V. RESULT ANALYSIS

In the context of the "Credibility Analysis", the crucial aspect lies in the model's ability to correctly predict or separate the classes and to correctly predict and assign topics. The "confusion matrix" serves in delineating the errors made by a classification model under consideration. The Precision and Recall values are computed for SVM classifier and depicted in Table 2. The outcome analysis of working with LDA and enhanced LDA are portrayed in Table 3. The outcomes demonstrate that with LDA more percentage of tweets got characterized into the classification "likely to be false" denoting vagueness in choice of apt and proper classification, whilst enhanced LDA gives much better results. However, these computed results may likely vary with the inputs provided, viz. the search keyword and the scale factor determining the count of the tweets intended to be retrieved.

TABLE -2 SVM PRECISION AND RECALL VALUES

	Classes		
	Positive	**Negative**	**Neutral**
Precision	89.2%	96.5%	22.22%
Recall	100%	66.6%	60.9%

TABLE -3 RESULTS OF LDA AND IMPROVED LDA

	% of Tweets in each category	
Degree Of Credibility	**LDA**	**Improved-LDA**
Almost certainly true	12.9%	70.9%
Almost certainly false	30.6%	8%
Likely to be false	56.45%	2.96%

VI. CONCLUSION AND FUTURE WORK

The paper strives to demonstrate the significance and role of Credibility Analysis in the milieu of the twitter data. To facilitate the analysis of the retrieved tweets, the opinion mining techniques viz. pre-constructed opinion lexicons and machine learning classifiers have been discussed from the implementation perspective, primarily emphasizing elaboration on Support Vector Machine classifier. Improved LDA has been utilized in order to associate each tweet with topic. Majority decision method is applied to derive credibility. Relevant to the context of discussion on the above mentioned information credibility analysis, the need to automate them and detail of the working procedure of the automated tool E-OPTIMA (**E**nhanced- **OP**inionated **T**weet **I**mplied **M**ining and **A**nalysis) ver. 2.0.0. have been explained and illustrated in a lucid manner. The paper culminates presenting discussion on the automated tool's ability to present results both graphically and in tabular format.

The future work entails remodeling the tool with better and improved credibility analysis techniques by incorporating tweets explicitly attributable to "working domain user expertise". The inclusion of such proficient information for credibility analysis is attributable to the fact that usually individuals' believability on the information increases if the source of information is justified by the creator skill. Another aspect of enhancement lies into fine tuning the tool with prime focus on emoticon oriented tweets and tweets attributed to inherent mixed languages like hinglish (a blend of Hindi and English) tweets. Apropos to these enhancements, the result analysis on the outcome produced by the improved version of the tool will evidently be subjected to relevant comparative analysis on progression/retrogression on the credibility aspect produced by the current vis-a-vis improved tool.

REFERENCES

[1] IBM official site, Beyond sentiment analysis: social data analytics [Online]. Available: http://www-01.ibm.com/software/ebusiness/jstart/socialdata/

[2] G. Vinodhini ,R. M. Chandrasekaran , "Sentiment Analysis and Opinion Mining: A Survey", in International Journal of Advanced Research in Computer Science and Software Engineering, Volume no. 2, Issue no. 6, 2012, pp. 282-292.

[3] S. Tan, Y. Li, H. Sun, Z. Guan, X. Yan, J. Bu, C. Chen, X. He, "Interpreting the Public Sentiment Variations on Twitter", in IEEE Transactions On Knowledge And Data Engineering, Volume no. 26, Issue no. 5, May 2014, pp. 1158-1170.

[4] Y. Ikegami, K. Kawai, Y. Namihira, S. Tsuruta, "Topic and Opinion Classification based Information Credibility Analysis on Twitter", in IEEE International Conference on Systems, Man, and Cybernetics, 2013, pp.4979-4681.

[5] M. Miyabe, A. Umejima, A. Nadamoto and E. Aramaki, "Proposal of Rumor Information Cloud based on Rumor-Correction Information" (In Japanese), RRDS4-019, 2011.

[6] F. Toriumi, K. Shinoda, G. Kaneyama, "Accuracy Evaluation of Demagogue Detection System using Social Media" (In Japanese), IPSJ Digital Practice, 3.3, pp. 201-208, 2012.

[7] D. Murdoch, "R-3.3.0 for Windows (32/64 bit)" [Online]. Available : https://cran.r-project.org/bin/windows/base/.
[8] S. Liu, X.Cheng, F. Li, "TASC:Topic-Adaptive Sentiment Classification on Dynamic Tweets", in Ieee Transactions On Knowledge And Data Engineering, Volume no. 27, Issue no. 6, June 2015, pp. 1041-1047
[9] Jayashri Khairnar*, Mayura Kinikar, "Machine Learning Algorithms for Opinion Mining and Sentiment Classification", June 2013.
[10] Timothy P. Jurka, "Package-Sentiment", 2012.