Real Time Analysis of Social Media Data to Understand People Emotions Towards National Parties

Santoshi Kuamri
Department Of CSE
M. S. Ramaiah University of Applied Sciences
Banaglore, India
santoshik29@gmail.com

Narendra babu C
Department Of CSE
M. S. Ramaiah University of Applied Sciences
Banaglore, India
narendrababu.c@gmail.com

Abstract—**Social media has given new way of communication technology for people to share their opinions, interest, sentiments. Huge amount of unstructured data is generated from social media like Facebook, twitter, LinkedIn, which is repository of useful insights. Analytics can be applied to extract various useful insights form this. Main objective of this paper is to extract the knowledge from large social media data, identify the people sentiments and behavior to make cognizant decisions. Effective sentiment analysis of people emotions is necessary for complex topics like politics, government, present trends and healthcare. These objective are achieved by real time retrieval of twitter data and perform sentiment analysis. It helps to classify emotions of the people are positive, negative, happy, sad or neutral towards national parties of India (BJP and INC). Considering people emotions, parties can identify their weakness and look after improvement to fulfill society needs and work towards satisfying people requirements. Using text mining and unsupervised lexical method classified tweets related to these parties to identify people emotions for the parties.**

Keywords— social media; data analytics; text mining; lexicon; unsupervised; sentiemnt analysis;

I. INTRODUCTION

In this big data era huge unstructured data generated continuously from social media such as Facebook, twitter, LinkedIn, blogs[1]. Social media allows user to easy way of sharing and communicating their opinion and emotions to the world in their own language or in the form of picture. Around 500 million people write their opinion and sentiments to the world on twitter every day and 6000 per second[2]. Since 2007 to till 2017 in small span of time, number of tweets increased 5000 to 500,000,000 tweets per day that is into 6 times more in 10 years. In future it may increase by 12 to 14 times more in next 10 years as many devices are connected to net.

Social media give freedom to people to express and share their views and opinion on present or past event, lifestyle, trends, government, healthcare, education, business, politics many more activities around us[3]. Large data generated form these social media in various format, various languages in high speed. It gives rise to many challenges in data analytics and computing area to find new value and extract knowledge from it and find sentiment and opinion on different issues and predict the future conditions[4][5].

Social media analytics is not only applying analytical and Machine Learning methods to identify and extracts knowledge to carry out sentiment analysis and opinion mining. On the other hand it is important to identify distinct domain for which gathering heterogeneous data form different place, different time, different format, mine and extract values and analyze from deferent perspectives, getting new values, identifying emotions and sentiments, type of events happening, identifying trends, change in trends and behavior[4].

Social media analytics is one of the challenging application among different big data analytical applications, as it large volume of unstructured data growing with high velocity from different locations of the world. To understand costumer requirements and opinions EBay.com uses two data warehouses 7.5 petabytes and 40PB Hadoop cluster. Amazon.com handles millions of back-end operations every day.

Fast growth of this social network motivated to do research and understand the social media users understand emotions. We have presented a new view to identify meaningful information from network based on people opinion on national parties of India. Developed an approach for mining communities to understand and analyze the network structure. This user interest based model can be used to identify society requirements and improve quality of governance for different location.

This paper aims to identify and retrieve valuable information and get knowledge from huge social media data and help national parties to understand people requirements and emotions and satisfy their needs. Specifically tweeter data related to Bharatiya Janata Party, hence forth referred as BJP and Indian National Congress, hence forth referred as INC is considered for emotion analysis. This can be achieved through following objectives. Extract real time data and pervious data from twitter. Clean and preprocess using text mining approach and calculate the frequency of words. Finally identify the behavior and people emotions towards parties by classifying the data into positive and negative polarity as well as detect eight different emotions of people for the parties using unsupervised lexicon method. This gives social media impact on new application such as government, politics, understudying society needs and improving in that directions to fulfill society needs.

The rest of the paper is organized as follows. Section II provides a related work on social media analytics for different applications along with politics and methods used for sentiment analysis of social data. Section III explains on proposed model for sentiment and emotion analysis of social data. Section IV provides results and discussion on sentiment analysis of real time data. Benefits of unsupervised analysis and future work are also discussed in this Section. Conclusion with recommendations for future work is presented in Section V.

II. Related Work

There are various applications of social media analytics such as link prediction, people behavior analysis, customer sentiment and requirement analysis [5], location based people behavior analysis, event detection and handling event so on. Following gives literature on past contributions to social media analytics in political related research areas.

Developing spatial based Bayesian model [6] aims to find political leaders, people are following using social network analysis. Applied this method to large data sample of twitter from US and five European countries. Many political activities using tweeter data in 2011 legislative elections of Spain and presidential elections of US is carried out in [7] by analyzing user behavior and representativeness towards political parties. Analysis shows most of the people are tweeting are from urban area and men also strong political inclinations.

Developed a tool sent meter for sentiment analysis of social data [8] for analyzing opinion data for the purpose of decision making. Tool uses a developed algorithm for estimating social campaign success rate. Using this tool identified location wise success rate of campaign in India called "Swachh Bharat Abhiyan"in 2014 using tweeter data. It achieved 84.47% of accuracy using manual tagging and unigram machine learning approach.

To identify and analyze political discussions on social media, [9] proposes a method to get deeper insights from political discussions. To implement a method for detecting political abuse and tracking them on social media [10] describes machine learning framework and detect viral spreading misinformation of politics. It combines content based feature extraction and topological method and gets 96% of accurate results on twitter data on 2010 US midterm elections.

Statistical sentiment model for sentiment analysis at real time on twitter data during 2012 US elections is proposed in [11]. It aims to identify real time political events and sentiment changes with time. It proposes generic model which can be applied to any domain like politics, movie sentiment analysis. Social media analytics has turn into progressively applicable political establishments and government area [12] [3] also in business and marketing [5][13].

Analytics on political data varies based on specific intentions and perspectives like topic based, exploratory based, actor based, event based or self-involved. Many text mining techniques are applied to identify topics and contents specifically content analysis of social media[14]. Content analysis is a technique used to identify and analyze valid information from text for particular context of research [14]. For large amount of social media data text mining and content analysis techniques are necessary for automated quantitative analysis. To answer various kind of queries these techniques are suitable such as identifying and modeling topics for classification of text.

Many different methods are originated such as logical methods includes dictionary based coding [15][16] and co-occurrence or word frequency analysis, many other new methods [17][18] have been implemented for automated text mining and social media analytics. There are various kind of supervised and unsupervised classification or learning methods [19][21] as well as semantic network analysis methods are applied for text mining and analysis for social data. Particularly statistical machine learning algorithms like support vector machine[18], Naïve Bayes and maximum entropy are became standard potential algorithms for text mining and classification [20]. Clustering algorithms k-means and hierarchical clustering algorithms are unsupervised learning techniques for document clustering[19].

Sentiment analysis of social media data performed by machine learning approach and dictionary based or corpus based approach as a two different methods. Most popular supervised machine learning methods for sentiment analysis is naïve Bayes and support vector machine which formulates sentiment analysis problem as learning and classification problems as positive, negative and neutral [21]. Recently several research work has been done on finding and classifying emotions like joy, sad, happy, and surprised. Unsupervised learning uses NLP techniques such as part of speech, bag of word, word frequency for learning and classification of text[4] to improve accuracy.

In dictionary based approach, dictionary with large list of word with their sentiment orientations is used to calculate and identify sentiment if text in the document.

Recently many advanced natural language processing methods are used for topic modeling to discover abstract topics from documents. Which includes specific statistical models such as Latent Dirichlet allocation (LDA) and Probabilistic latent semantic analysis (PLSA) [22] along with nonnegative matrix factorization or singular value decomposition.

III. Proposed Model

In proposed approach plans to determine people trust and emotions towards national parties of India based on emotion analysis and polarity classification using unsupervised lexicon method. Here we make use lexicon as an unsupervised method for extracting emotion and sentiments from tweeter data by following steps as shown in Fig. 1. Process carried out in each steps are described as following.

1. Data collection: Tweeter data is extracted and collected using twitter API. 10000 unstructured tweets are extracted at real time based on the concept BJP and INC.

2. Preprocessing: tweets contains unwanted features like http, @ , #, www, url, ., “” , so on. All unwanted data are cleaned and removed to extracted useful text by applying text mining techniques.

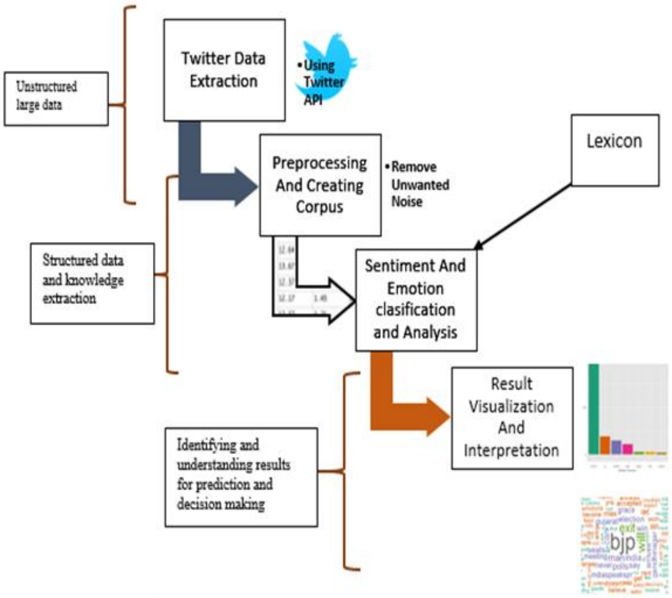
Unstructured large data
Twitter Data Extraction
•Using Twitter API
Preprocessing And Creating Corpus
•Remove Unwanted Noise
Lexicon
Structured data and knowledge extraction
Sentiment And Emotion clasification and Analysis
Result Visualization And Interpretation
Identifying and understanding results for prediction and decision making
bjp

Fig. 1. Proposed System For Emotion Anaysis

3. Feature selection: In this step data is normalized and converted to lower case letters. Entire document is tokenized into set of words or corpus. Subsequently stemming is done to decrease relevant tokens into single type.

4. Feature weighting: Determines the frequency of words in a collection of documents. In which rows correspond to documents in the collection and columns correspond to terms. Using Natural language processing methods determine importance of a word in document or corpus of words.

   Using importance of words in the document created a word cloud where each word size indicates the importance of that word in that topic.

5. Emotion classification: in this step we have used unsupervised lexicon based approach as it is less sensitive to change of topic for concept level sentiment analysis and it can be applied to heterogeneous applications. NRC lexicon method is used to associate the emotions and sentiment of collected tweets and calculate emotion level.

6. Emotion interpretations: graphical representation of results is easy to visualize and understand the results. In this step 10 different kind of emotions are depicted to understand the emotions from collected tweets.

IV. Results and Discussion

A. *Extracting Data From Twitter*

Extracted 10000 tweets from JAN to APRIL 2017 at real time using tweeter API connecting it to R-Studio. Twitter data is considered for analysis because of its easy accessibility and tweets of the people are spoken globally instead of in a group like Facebook, LinkedIn.

Next is to identify and generate feature set from unstructured data from tweets. To generate feature set we extract text from comments and remove noise and unwanted data. The extracted text is then split into words signifying feature that contains hidden meaning in it. So that we can count number of times the feature or word occurs in positive and negative comment.

B. *Determine the Frequency of Words*

A mathematical model term-document matrix (TDM) is used which determines the frequency of words that occur in a collection of documents. The rows correspond to documents in the collection and columns correspond to terms. This model uses one of the Natural language processing method TF-IDF (term frequency–inverse document frequency) to determine importance of a word in document or corpus of words. TF-IDF is used to create a sentiment word list from the text document under vector space model. TF calculates how frequently a term occurs in a document and IDF measures how important a term by using following equations.

TF (t) = (Number of times term t appears in a document) / (Total number of terms in the document).

IDF (t) = log_e(Total number of documents / Number of documents with term t in it).

TDM is constructed using INC and BJP related twitter data

TDM for INC: TD Matrix of INC tweet data contains 5101 terms in 5000 documents, 66199 non sparse entries and 25438801 sparse entries. Maximum term length obtained is 29 with 100% sparcity.

TDM for BJP: TD Matrix of BJP tweet data contains 5029 terms in 5000 documents, 67456 non sparse entries and 25077544 sparse entries. Maximum term length obtained is 27 with 100% sparcity.

Analysis of above data is done in two steps: one is using Text Mining (tm) package. Step one provides various text mining methods for extracting knowledge from unstructured data. In step two syuzhet package is used for sentiment analysis of tweets. In syuzhet package, NRC lexicon algorithm is used. Key advantage of using it is, it associates the emotions to words to analyze sentiments

C. *Dimensionality Reduction*

The data obtained after preprocessing contains a large set of words, however it is still multidimensional. To reduce understanding difficulties it becomes essential to decrease dimensions**.** Latent Semantic Analysis (LSA) is used to analyze relationship between set of words and documents which makes meaningful sentences in the document. It assumes the words in the same piece of the document will have closer meaning. TDM is essentially a very sparse matrix (99% sparseness is very common). So to reduce sparseness a matrix is constructed containing word count per paragraph where row represents unique word and column represents paragraph.

INC: After reducing sparcity, TDM of INC tweet data contains 47 frequent terms in 5000 document, 27910 non sparse entries and 207090 sparse entries. Maximum term length obtained is 10 with 88% sparcity.

BJP: After removing sparcity, TDM of BJP tweet data contains 47 terms in 5000 documents, 24092 non sparse entries and 210908 sparse entries. Maximum term length obtained is 18 with 90% sparcity.

D. Word Cloud Construction

Using frequency and importance of words in a document or topic a word cloud is constructed where each word size indicates the importance of that word in that topic as shown in Fig. 2 and Fig. 3.

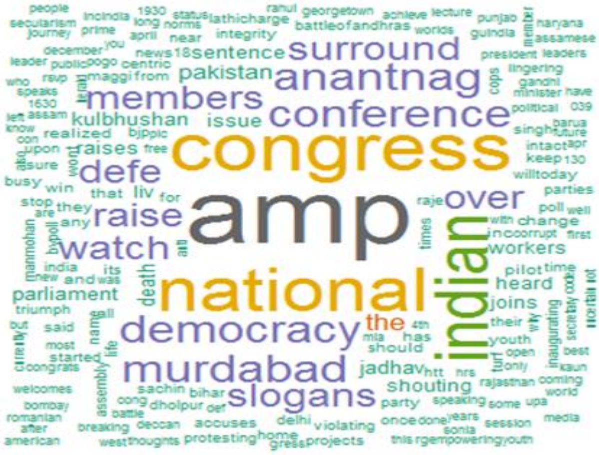

Fig. 2. Word Cloud on INC Related Data

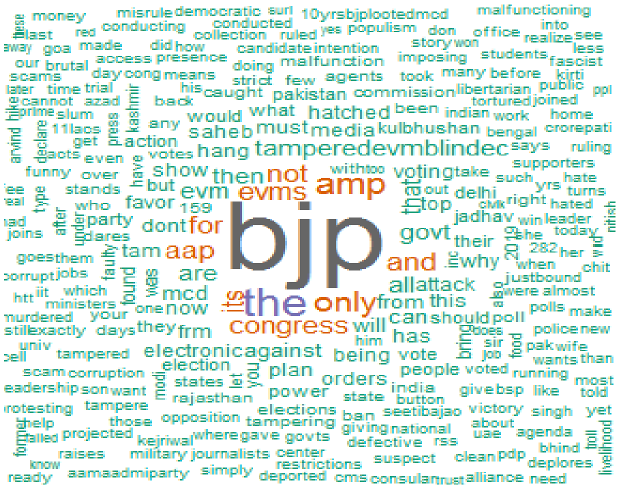

Fig. 3. Word Cloud on BJP Related Data

People used more number of words such as bjp, Indian, congress, amp, democracy, slogan, and shouting. The surrounding words are of lesser importance and lesser the frequency count the smaller is the font size of the words. Word cloud construct depicts people have compared INC and BJP and similar other parties.

E. Sentiment and emotion analysis

It is important to understand people opinion, expectations and emotions towards the national political parties. This can be achieved by analyzing the social media data.

Sentiment analysis is carried out aiming to determine the attitude of a speaker or people with respect to topic or the overall contextual polarity of a document. A basic task in sentiment analysis is to classify the polarity of a given text at the document or sentence level as positive, negative, or neutral. Further advanced sentiment classification is done to identify emotional states such as “angry,” “sadness,” and “joy”.

In this step syuzhet is used for sentiment extraction from the text and depict the sentiment curve for the given text. It includes four sentiment lexicons such as NRC, AFFIN, BING, and Stanford for sentiment and emotion analysis of text. Proposed approach is implemented using NRC sentiment dictionary to calculate positive and negative sentiments along with eight different types of emotions such as "anger", "anticipation", "disgust", "fear", "joy", "sadness", "surprise", "trust” valence from the 10000 tweets.

TABLE I. EMOTION COUNT FOR INC

SI No	Emotions for INC	Count
1	anger	1948
2	anticipation	1346
3	disgust	4746
4	fear	1838
5	joy	1104
6	sadness	1041
7	surprise	1698
8	trust	6937
9	negative	3146
10	positive	4160

TABLE II. EMOTION COUNT FOR BJP

SI No	Emotions for BJP	Count
1	anger	1781
2	anticipation	1980
3	disgust	1438
4	fear	1703
5	joy	1312
6	sadness	1136
7	surprise	787
8	trust	4066

9	negative	3428
10	positive	3951

Sentiments count of 10000 tweets are generated for INC and BJP respectively as shown in Table I and Table II. Fig 4 and Fig 5 are the graphical representation of people emotions towards the political parties BJP and INC. Where horizontal axis represents ten different emotion count as anger, anticipation, disgusting, fear, joy, negative, positive, sadness, surprise, trust and vertical axis represents the total sentiment counts.

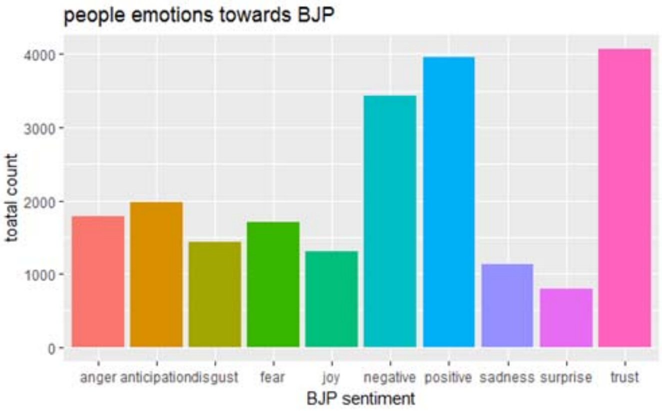

Fig. 4. Emotion vizualization on BJP sentiment count

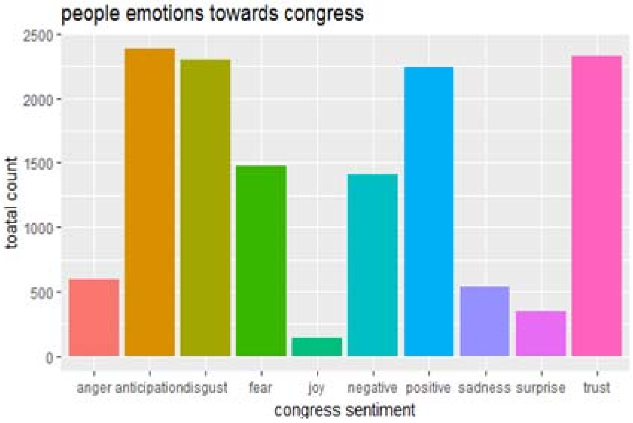

Fig. 5. Emotion vizualization on INC sentiment count

Visual representation makes it easier to understand emotions of people towards political parties. Meanwhile it is important for political parties to identify and understand people and society requirements and work towards solving society issues.

Accuracy of this model is low compare to machine learning as it uses only unigram dictionary. This work can be extended as a future work by combining naive Bayes and lexicon methods to enhance accuracy for real time heterogeneous data then validating the model using confusion matrix.

Advantages of the proposed approach are: Scalable, Real time and Simple, instantaneous results. Making use of unsupervised lexicon method is scalable for heterogeneous application to get quick results in Real time and identify emotions and analyze data. Data in social media like tweeter and Facebook, gets generated on several situations, events and trends with change in time on a scale of seconds or minutes. Earlier discussions, forums in social media span for a day to a week or a month. But now a day's people trends and discussion changes in seconds based on type of event and topic. However events or incidents gets related politics and greater participation of people in social media. Therefore it is necessary to analyze the real time data to make better decisions. Hence this method is suitable for identifying and analyzing people emotions at real time.

This work can be extended in several aspects. We basically focused on people emotion towards the political parties, considering specifically the larger national parties. Future work will be to expand the real time analysis across several domains such as Security, Healthcare, Business and many other recent trends where the real time responses are critical. Further the second line of future work is toward improving the accuracy by combining learning and lexicon method at real time. It would be more challenging to accomplish higher scalability, accuracy and performance at real time analysis of large social media data.

V. Conclusion

This paper provides an approach on collecting, storing, processing, analyzing and summarizing political related social media data in real time and identifying people opinion and emotions towards particular political party. Focusing mainly on identifying people emotions and sentiments towards two national parties of India (BJP and INC) by extracting the knowledge from tweeter data and sentiment analysis on real time and past tweets. Proposed work involves retrieving the data from tweeter using tweeter API, normalization of data using text mining techniques, applying NLP methods to calculate frequency of word and document, construct word cloud. Finally calculate polarity and emotions using unsupervised lexicon based approach. Even though this paper aims towards implementation of methods for social media analytics to extract knowledge, it is important for political parties to identify, understand people and society requirements then work towards solving society issues. Extension of this work is to combine machine learning and lexicon method for heterogeneous applications sentiment analysis, opinion mining for better decision making and prediction.

References

[1] Bello-orgaz, G., Jung, J. J., & Camacho, D. (2016). Social big data: Recent achievements and new challenges. Information Fusion, 28(August), 45–59. doi:10.1016/j.inffus.2015.08.005

[2] Twitter Usage Statistics - Internet Live Stats. (n.d.). Retrieved April 14, 2017, from http://www.internetlivestats.com/twitter-statistics/

[3] Kavanaugh, A. L., Fox, E. A., Sheetz, S. D., Yang, S., Li, L. T., Shoemaker, D. J., … Xie, L. (2012). Social media use by government: From the routine to the critical. Government Information Quarterly, 29(4), 480–491. doi:10.1016/j.giq.2012.06.002

[4] Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies. Retrieved from

http://www.morganclaypool.com/doi/abs/10.2200/s00416ed1v01y201204hlt016

[5] Stieglitz, S., Dang-Xuan, L., Bruns, A., & Neuberger, C. (2014). Socialmedia analytics. Business and Information Systems Engineering, 6(2), 89–96. doi:10.1007/s12599-014-0315-7

[6] Barberá, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. Political Analysis. Retrieved from https://www.cambridge.org/core/journals/political-analysis/article/div-classtitlebirds-of-the-same-feather-tweet-together-bayesian-ideal-point-estimation-using-twitter-datadiv/4FAF590D76248A33C2F7D03D1ECA9E70

[7] Barbera, P., & Rivero, G. (2015). Understanding the Political Representativeness of Twitter Users. Social Science Computer Review, 33(6), 712–729. doi:10.1177/0894439314558836

[8] Tayal, D. K., & Yadav, S. K. (2016). Sentiment analysis on social campaign "Swachh Bharat Abhiyan" using unigram method. AI & SOCIETY. doi:10.1007/s00146-016-0672-5

[9] Stieglitz, S., & Dang-Xuan, L. (2013). Social media and political communication: a social media analytics framework. Social Network Analysis and Mining, 3(4), 1277–1291. doi:10.1007/s13278-012-0079-3

[10] Ratkiewicz, J., Conover, M., Meiss, M., & Gonçalves, B. (2011). Detecting and Tracking Political Abuse in Social Media. ICWSM. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2850/3274/

[11] Wang, H., Can, D., Kazemzadeh, A., Bar, F., & Narayanan, S. (2012). A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, (July), 115–120. doi:10.1145/1935826.1935854

[12] Stieglitz, S., Brockmann, T., & Xuan, L. D. (2012). Usage Of Social Media For Political Communication. PACIS 2012 Proceedings. Retrieved from http://aisel.aisnet.org/pacis2012/22

[13] Larson, K., & Watson, R. (2011). The value of social media: toward measuring social media strategies. Retrieved from http://aisel.aisnet.org/icis2011/proceedings/onlinecommunity/10/

[14] Krippendorff, K. (2004). Content analysis: An introduction to its methodology. Retrieved from https://books.google.co.in/books?hl=en&lr=&id=q657o3M3C8cC&oi=fnd&pg=PA3&dq=content+analysis+an+introduction+to+its+methodology&ots=bLhdw1Kdx0&sig=yCzopN0NW95Yp2y83cyJxcP-Bo8

[15] Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., & Liu, B. (2011). Sentiment Analysis Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis.

[16] Taboada, M., Brooke, J., & Voll, K. (2011). Lexicon-Based Methods for Sentiment Analysis, (September 2010).

[17] Hillard, D., Purpura, S., & Wilkerson, J. (2007). An Active Learning Framework for Classifying Political Text. Annual Meeting of the Midwest Political Science Association, (429452), 1–31.

[18] Bifet, A., & Frank, E. (2010). Sentiment Knowledge Discovery in Twitter Streaming Data (pp. 1–15). doi:10.1007/978-3-642-16184-1_1

[19] Montejo-Ráez, A., Martínez-Cámara, E., Martín-Valdivia, M. T., & Ur Na-López, L. A. (2012). Random Walk Weighting over SentiWordNet for Sentiment Polarity Detection on Twitter, 3–10. Retrieved from http://www.aclweb.org/anthology/W12-3703

[20] Scharkow, M. (2013). Thematic content analysis using supervised machine learning: An empirical evaluation using German online news. Quality & Quantity, 47(2), 761–773. doi:10.1007/s11135-011-9545-7

[21] Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. Retrieved from http://arxiv.org/abs/cs/0205070

[22] Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. Journal of machine Learning research. Retrieved from http://www.jmlr.org/papers/v3/blei03a.html