International Journal For Technological Research In Engineering
Volume 4, Issue 8, April-2017 *ISSN (Online): 2347 - 4718*

OPINION MINING TO PREDICT ELECTION RESULTS

Brahmbhatt Akash R[1], Risha Tiwari[2]
[2]Asst.Professor(CE), [1,2]HGCE,vahelal.

Abstract: ***An expeditious development of the web Application has created countless opportunities for the public to vocalize their Opinions. Web Applications like Social Media, Blogs, News Portal, E-commerce Sites etc. has created huge buzz amongst people. Now a days amongst all above web applications Social media has become the mirror of the society. Starting from giving opinions about the movies or checking popularity of any celebrity, or discussion about political parties during elections or predicting box-office results or predicating election results or starting of new trend all things become easier because of the gaining popularity of Social media. Amongst all Social Media twitter has huge tie up with all the above mentioned stuffs. Now all this discussions, or giving an opinion or giving star credits or reviews generate huge amount of user generated content, to analyze and summarize this content is the job of natural language processing and machine learning techniques. For example opinion mining or Sentiment Analysis, text mining, graph mining etc. In this paper all opinion mining techniques and supervised machine learning techniques will be discussed which will be useful to predict election results from the tweets. And we will also compare weka results with our own implemented results. This paper contains introduction following by related work, opinion mining techniques, purposed method and results and conclusion.***

Index Terms: Social media, Opinion Mining, Twitter, Election Results ,Emotions ,Political Opinion, Sentiment Analysis, Supervised Machine learning.

I. INTRODUCTION

Online Social Media don't need any introduction, it's the newest entertainment or we can say time pass medium for this generation and perhaps for all the future generations. Social media has become a major part of most people's lives and it's been changing an aspect of their lives in many ways since last two or three years. It has become one of the key communication medium used over an internet. People publish their opinions on variety of topics and discuss latest trends, post their daily life activities. Social medias like Twitter, My space, Facebook, Instagram are at the peak amongst public. Though amongst all the above mediums Twitter is favorite amongst research analysts. Twitter is an online social networking and micro blogging site which allows user to send and read messages known as tweets, which has 140 characters limit. It has millions of active users. User can forward or retweet other user's tweets to his or hers followers. A user can also use @mentions to refer a particular user. In addition now a days twitter is a new medium for political debates and movie review discussions. Millions of tweets are being generated and forwarded before the day of an election. Many researchers extract tweets from twitter and analyze it using machine learning techniques to predict future trends and results which are going to be discussed here. Opinions are often expressed in large texts though natural language processing techniques like text-processing can be used to find that opinions.

In Current scenario there are some methods which has been established to categorize opinions as positive or negative. Methods like un-supervised learning and supervised learning can be used to establish above task.

In these paper that methods are discussed in brief.

II. BACKGROUND AND RELATED WORK

Many Researchers provided research work in social media and sentiment analysis. Sentiment analysis is a text processing technique to derive an opinion or mood based on the terms used in a large sentence. A huge number of researchers have focused on generating statistical inference from social media data using sentiment analysis techniques.

Gautami Tripathi and Naganna s[1] presented a survey on sentiment analysis and the related techniques. They also discussed the application areas and challenges for sentiment analysis with insight into the past researches. Yulan He, Hassan Saif, Zhongyu Wer, kam-Fai Wong[2] proposed a statistical model for sentiment detection with side information. and they map the public opinion in twitter with the actual offline sentiment in real world. Diego Tumitan and Karin Becker [3] demonstrated Tracking Sentiment Evolution on user-generated content: A case study on the Brazilian Political Scene. They used an approach to identify and classify sentiment in news comments written in Portuguese Language. G.Angulakshmi & Dr. R. Manickar Chezian [4] analyzed all the Opinion mining Techniques and tools. Nishantha Medagoda, Subana Shanmuganathan, Jacqueline Whalley[5] investigated opinion mining and sentiment classification studies in three non-english languages to find the classification methods and efficiency of each algorithm. Muhammad Asif Razzaq, Ali Mustafa Qamar, Hafiz Syed Muhammad Bilal[6]depicted that social media content can be used as an effective indicator for capturing political behaviours of different parties positive, negative, neutral behavior of the party followers as well as party's campaign impact can be predicted from the analysis. Antoine Boutet, Hyoungshick Kim and Eiko Yoneki [7]showed that the experimental results showed that the number of twitter messages referring to a particular political party-achieved about 86%classifiaction accuracy without any training phase.

III. OPINION MINING

Opinions of others highly influence the human behavior and are central to almost all decision making activities[1]. The major part of our information gathering process is to find out

what others think[1].Opinion mining also named as a sentiment analysis. It's a way to find user's view on particular topic. That topic can be any a movie or politics or product etc. Opinion mining is a part of a text mining and web mining. The goal of Opinion Mining is to make computer able to recognize and express emotions [4]. A thought, view, or attitude based on emotion instead of reason is called sentiment[4]. Opinions can be of different types- direct, indirect and comparative[1].

- Direct opinion expresses the direct views about an object. For example, "President Obama is doing a great job.[1]"
- Indirect opinion indirectly expresses the views about the object. For example, "I had a headache after watching the movie" gives a negative review about the movie[1].
- Comparative opinions describe the likes and dislikes of the opinion holder about an object over the other. For example, "Samsung mobiles are better than Nokia.[1]"

The task of opinion mining can be performed at three different levels namely, document level, sentence level and aspect level[1].At document level the entire document is classified as positive, negative or neutral[1]. This level of analysis assumes that the entire document expresses opinions on single entity (object)[1]. It is not applicable to documents evaluating multiple objects. At sentence level we analyze each sentence and determine its polarity (positive, negative or neutral)[1].

At aspect level we perform a finer analysis by identifying and extracting the product features from the source data[1].

3.1 Workflow of opinion mining

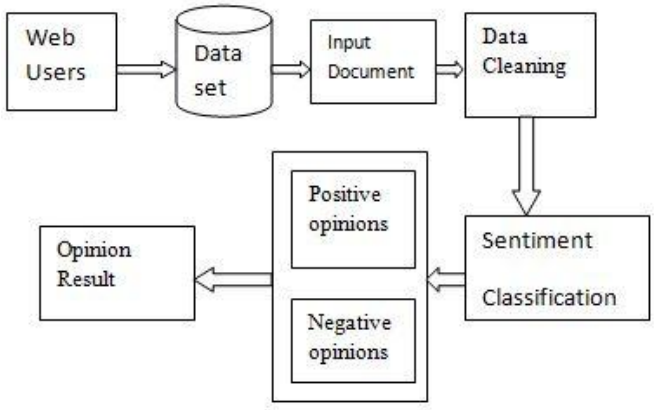

Figure 1:Workflow of Opinion mining

As Shown in Figure 1 Opinion mining start by crawling user generated data either in the form of comments or tweets or blogs or reviews etc. After that the data will be taken in input text document like csv file to perform preprocessing tasks. In preprocessing tokenization of the text, and stemming take places. Then sentiment classification task take places, in that either un-supervised or supervised learning method is used to classify the data in to positive and negative opinions. In that step given a review document D and a predefined categories set c={positive, negative} Sentiment classification is to classify each document, with a label expressed into two forms namely positive and negative[4].

The last step checks the impact of opinionated text on the given document,means whether the document has positive polarity or negative.

3.2 Opinion mining Techniques

To judge the polarity of the given document can be done by machine learning algorithms[5]. The choice of the specific learning algorithm used is a critical step[5]. The methods of determining the semantic orientation used for identifying the polarity of the sentence are categorized into two approaches: supervised and unsupervised classification techniques[5].

3.2.1 Un-Supervised lexicon based Technique

Work: Lexicon is composed of a set of positive and negative words, used to score the opinion sentences either positive, negative or neutral. This approach is very popular and requires a scoring function to score every sentence according to the existence of positive or negative words[1].

Limitations: It's hard to handle the negation based sentences.

3.2.2 Supervised Classification Techniques[5]

Work: Supervised classification algorithm is one of the learning algorithms most frequently used in text classification systems. In supervised classification two sets of opinion sets are required namely, training and testing data sets. The training data set is used to train the classifier to learn the variation of the characteristics of the sentence or document and the test data is used to validate the performance of the classification algorithm. The supervised machine learning techniques, such as Naïve Bayes, support vector machines (SVM) and maximum entropy, Artificial neural network(ANN) are the most popular ones and they have been proven to be the most successful in sentiment classification[5].

Limitations: A supervised learning algorithm generally builds a classification model on a large annotated corpus.It's accuracy is mainly based on quality of the annotation, and usually the training process will take a long time.Besides that, when we apply algorithm to another domain,the result is usually not good

IV. MACHINE LEARNING SUPERVISED CLASSIFICATION IN BRIEF

In this section, some Supervised Classification methods will be discussed. And which will be used to predict the class of sentiment.

This classification algorithm requires two set:First is Training set: The training data set is used to train the classifier to learn the variation of the characteristics of the sentence or document[5].Second is Testing set: the test data is used to validate the performance of the classification algorithm[5].The supervised machine learning techniques, such as Naïve Bayes, support vector machines (SVM) and maximum entropy, are the most popular ones and they have been proven to be the most successful in sentiment classification[5].

Naïve bayes Classification:

Naïve Bayes algorithm is the most widely used and it is a simple but effective supervised classification method. The basic idea of the method is to estimate the probabilities of sentiment (either positive or negative) for the given opinion

using the joint probabilities of a set of words in a given category. The method is totally dependent on the naïve assumption of word independence[5].Naive bayes works fast for the training phase.Here,in equation class c* is assigned to document d.

$$c^* = argMax_C P_{NB}(C|D) \qquad \text{Eq1}$$

$$P_{NB}(C|D) = \frac{(P(c)\sum_{i=1}^{m} P(f|c)^{n,(d)})}{P(d)} \qquad \text{Eq-(2)}$$

Where, m is the no of features and f is the feature vector. Consider a training method consisting of a relative-frequency estimation P(c) and P (f | c)[8]. Despite its simplicity and the fact that its conditional independence assumption clearly does not hold in real-world situations, Naive Bayes-based text categorization still tends to perform surprisingly well; indeed, Naive Bayes is optimal for certain problem classes with highly dependent features[8].

Maximum Entropy:

Maximum Entropy (ME) classification is yet another technique, which has proven effective in a number of natural language processing applications. Sometimes, it outperforms Naive Bayes at standard text classification. Its estimate Of P(c | d) takes the exponential form as in Eq. (3)[8].

$$P_{ME}(C|D) = \frac{1}{Z(d)} exp\left(\sum_i \lambda_{i,c} F_{i,c}(d,c)\right) \text{ Eq.(3)}$$

Where, Z (d) is a normalization function. Fi, c is a feature/class function for feature fi and class c, as in Eq. (4),

$$F_{i,c}(d,c') = \begin{cases} 1 & n_i(d) > 0 \; and \; c' = c \\ 0 & otherwise \end{cases} \text{ Eq.(4)}$$

For instance, a particular feature/class function might fire if and only if the bigram "still hate" appears and the document's sentiment is hypothesized to be negative. Importantly, unlike Naive Bayes, Maximum Entropy makes no assumptions about the relationships between features and so might potentially perform better when conditional independence assumptions are not met[8].

Artificial Neural network[9]

Taking the nature of brain as model artificial neural network learn recognizes from experience. ANN is usually presented as systems of interconnected "neurons" that can compute values from inputs by feeding information through the network. For example, in a neural network for handwriting recognition, a set of input neurons may be activated by the pixels of an input image representing a letter or digit. The activations of these neurons are then passed on, weighted and transformed by some function determined by the network's designer, to other neurons, etc., until finally an output neuron is activated that determines which character was read. There are three types of learning in ANN, namely supervised, unsupervised and reinforcement learning. Perhaps the greatest advantage of ANNs is their ability to be used as an arbitrary function approximation mechanism that 'learns' from observed data. However, using them is not so straightforward, and a relatively good understanding of the underlying theory is essentiall[9].

Support Vector Machine[8]

Support vector machine is the best binary classification method[4].SVM is a non- probabilistic classification technique that looks for a hyper plane with the maximum margin between positive and negative examples of the training opinions[4].It has been proven that most probably svm provides better accuracy compared to another classification techniques.

How SVM Works?:

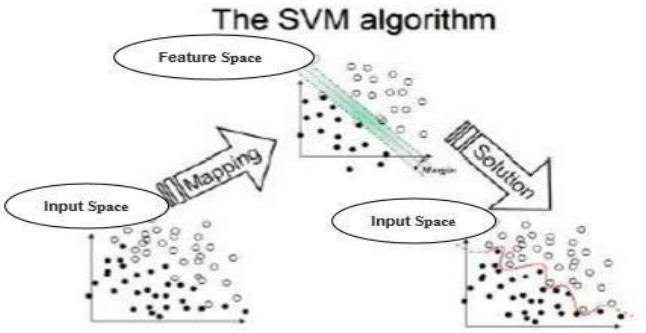

Figure 2:SVM Process Flow[8]

In Figure 6, data are input in an input space that cannot be separated with a linear hyperplane. To separate the data linearly, points are map to a feature space using a kernel method. Once the data in the feature space are separate, the linear hyperplane gets map back to the input space and it is shown as a curvy non-linear hyperplane. This process is what makes SVM amazing. The SVM's algorithm first starts learning from data that has already been classified, which is represented in numerical labels (e.g. 1, 2, 3, etc.) with each number representing a category. SVM then groups the data with the same label in each convex hull. From there, it determines where the hyperplane is by calculating the closest points between the convex hulls (Bennett, K. P., & Campbell, C., 2000). Once SVM determines the points that are closest to each other, it calculates the hyperplane, which is a plane that separates the labels[8]. When We use SVM with Unigram It gives More accuracy then it gives in Naïve bayes and Maximum entropy classification methods.

V. METHODOLOGY

5.1.1 There are Two Ways for Data Extraction:

Now, here our proposed method is to perform opinion mining to predict election results.

There are two ways by which we can perform this task:

First Method is to crawl data from the Social media and then after cleaning up that data perform Feature Extraction method then apply some classification techniques and at end it will give results in terms of % .

Second method is to fetch content from the E-news papers and then perform all the opinion mining techniques. We are going to perform this task on Twitter data. Because as we mentioned earlier that twitter provides user to write tweets in 140 characters and there are also some techniques by which we can fetch tweets easily compared to fetching data from E-news papers. From The Above mentioned Framework and Proposed Methods Of Application we have chosen Twitter To Perform Our Task.

Why Twitter????

- As mentioned above that twitter provides easy way to access it's data and it's easy to classify tweets so for that reason we have Chosen Twitter as our propose Application.

- On the Other hand it's hard to extract whole paragraphs related to politics from E-news Papers or Blogs.
- The Only Disadvantage to apply opinion mining on twitter data is that it's difficult to maintain "Sarcastic Comments".

5.2.2 Proposed Methodology: Opinion Mining To predict Election Results From Twitter Data:

How it Will Work?

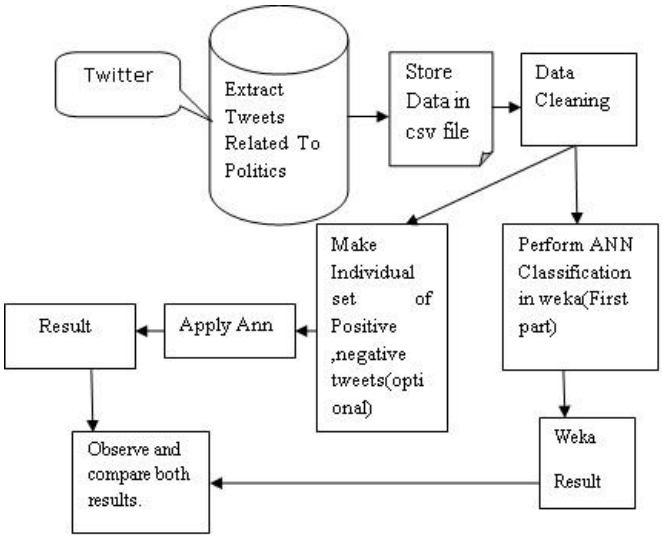

Figure 3 :Proposed Model To Predict Election Results.

Step 1:Data Extraction

- As, depicted above we are going to perform this task of predicting election results on twitter. So, Twitter provides various APIS to access it's data. Using that APIS we can fetch tweets from twitter.
- Here We are going to use Python's tweepy packages and twitter's API to crawl tweets.
- Data Crawling will be done by using @mentions and #tags related to political parties or political events.
- Here due to lack of decoding tools we are going to use banchmarke (universal) data of Us Election which is available on github.

Step 2:Data cleaning and store it in csv file

- So,As mentioned above that we are going to use benchmark data which is already in csv format.
- Here,java code will be used to tokenize the tweets in unigram.
- In tokenization it will remove comma seprated lines and will convert each line in to set of characters which is token.

Step 3:Perform Labeling using Unigram

- In this step that tokens which were generated in step 2 will be used.
- Each token will be work as a single character which called as unigram.
- Now using java code each unigram will be checked for it's weights.

Weights means weather it's positive polarity is high or negative.

Word	Positive_ weight	Negative Weight	Total
Official	7370000	1030000	6340 000
Proudly	3630000	3010000	6200 000
Represent	1680000	4370000	- 2690 000
I repress	7370000	1030000	6340 000

Table 1:Weighted keywords.

It will use each keyword of a record or tweet and check in google with two combination (Excellent and poor).

Ex:

Now suppose we have a word "Girl" So this process will combine Excellent+Girl and check for it's weight and Poor+Girl And Check for it's Weight.

Now this two weight will be subtracted which will give the total weight of the word"Girl".Which is a actual label of that word that it's actually positive or negative.

So this records will be stored in ".txt" file which we will use for further process.

Step 4:Make individual set of positive and negative tweets(Optional):

In this set we will classify tweets based on their weights which is stored in .txt file in two classes positive and negative tweets.

It's not compulsory we can directly use ".txt "file to apply classification algorithm without performing this step.

Step 5:Apply Ann(Artificial Neural Network)

Now in this step an actual work take places which is classifying tweets into positive and negative classes according to party names and get the result that who will win election based on no.of positive tweets.

Here, we will give labeled word as an input which will be multiplied with weight factor.

Weight factor will be decided by us, after multiplying each input with weight factor Ann will put them in one of the class either positive or negative.

Here 1st (It's like training phase)we will train a classifier that will put tweets in either of class. Then in next step it automatically put them in that classes.

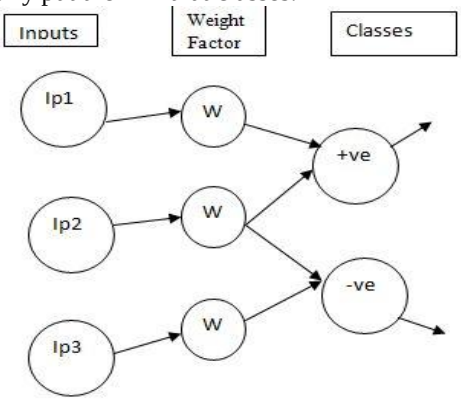

Figure 4 :Ann Workflow

Step 6:Examine the Result
At last Will have to Examine all the positive tweet counts of related party. And that party whose positive tweet will be more compared to other one will win that election. And we will also apply ann using weka on same dataset. and we will compare both results. Here we have graphs.
And then we will display that result in form of graphs.
1)Weka Results.

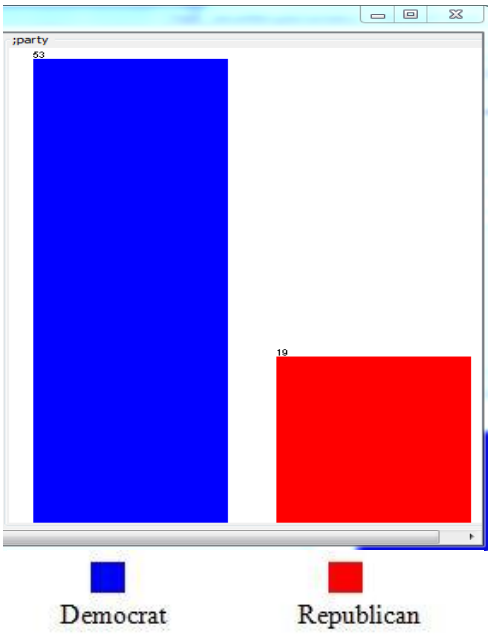

Figure 5:Weka result

2)ANN Results:

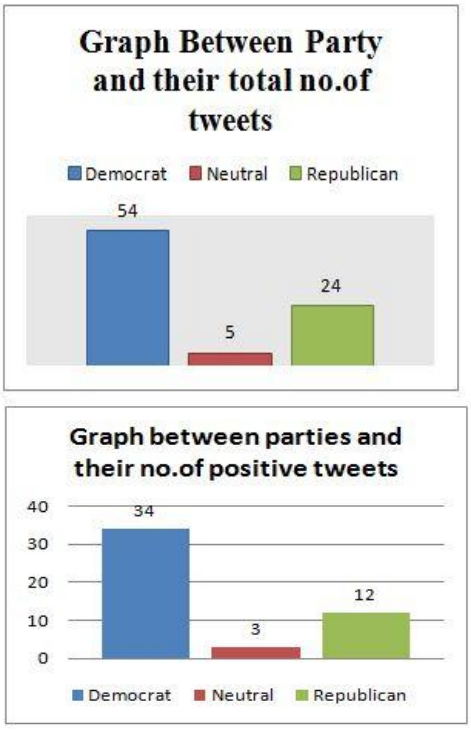

Figure 6 & 7 ANN Results

From Graph we can show that Our own method gives more accurate results than weka's results. Weka result only shows tweets related to party, we cant predict which party wins or lose from it. and in our results it shows total positive tweets according to party name from that positive tweets we can show that party which have higher no of tweets wins the election.

VI. CONCLUSION AND FUTURE WORK

To Conclude, This paper discusses all techniques of opinion mining, It also depicts all the supervised learning algorithm which is useful for classification in opinion mining. Here, It also discusses proposed method and algorithm. In Future we can use this model to predict movie collections, stock market etc.

REFERENCES

[1] Gautami Tripathi and Naganna S," Opinion Mining: A Review", International Journal of Information & Computation Technology.ISSN 0974-2239 Volume 4, Number 16 (2014), pp. 1625-1635 © International Research Publications House.

[2] Yulan He,Hassan Saif,Zhongyu Wei,Kam-Fai Wong"Quantising Opinions for Political Tweets Analysis"www.researchgate.net/publication/233934221,July-2015.

[3] Diego Tumitan, Karin Becker," Tracking Sentiment Evolution on User-Generated Content:A Case Study on the Brazilian Political Scene", Instituto de Informatica - Universidade Federal do Rio Grande do Sul, Brazil-2013.

[4] G.Angulakshmi1, Dr. R. Manicka Chezin," An Analysis on Opinion Mining: Techniques and Tools" ,Research Scholar, Department of Computer Science, Nallamuthu Gounder Mahalingam College, Pollachi, India, International Journal of Advanced Research in Computer and Communication Engineering Vol.3, Issue 7,July 2014.

[5] Nishantha Medagoda, Subana Shanmuganathan, Jacqueline Whalley," A Comparative Analysis of Opinion Mining and Sentiment Classification in non- English Languages ",Auckland University of Technology-2013.

[6] Razzaq, M.A. Sch. of Electr. Eng. & Comput. Sci. (SEECS), Nat. Univ. of Sci. & Technol.(NUST), Islamabad, Pakistan Qamar, A.M. ; Bilal, H.S.M. "Prediction and analysis of Pakistan election 2013 based on sentiment analysis", Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on- 17-20 Aug. 2014.

[7] Antoine Boutet INRIA Rennes Bretagne Atlantique France , Hyoungshick Kim and Eiko Yoneki Computer Laboratory, University of Cambridge UK," What's in Your Tweets? I Know Who You Supported in the UK 2010 General Election"2012.

[8] Jayashri Khairnar,Mayura kinikar, Department of Computer Engineering, Pune University, NIT Academy of Engineering,Pune," Machine Learning

Algorithms for Opinion Mining and Sentiment Classification",International Journal of Scientific and Research Publications,Volume 3,Issue 6,June 2013.

[9] Vikrant Hole,Mukta Taklikar" A Survey on Sentiment Analysis And Summarization For Prediction", International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 3 Issue 12 December 2014, Page No. 9503-9506.

[10] Eric Sanders CLS/CLST, Radboud University Nijmegen , Antal van den Bosch CLS, Radboud University Nijmegen Erasmusplein 1 6525 HT Nijmegen" Relating Political Party Mentions on Twitter with Polls and Election Results".

[11] Peter D. Turney Institute for Information Technology National Research Council of Canada Ottawa, Ontario, Canada, K1A 0R6 peter.turney@nrc.ca," Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews", Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 417-424.

[12] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, Noah A. Smith," From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series" International AAAI Conference on Weblogs and Social Media, Washington, DC, May 2010.

[13] Bo Pang and Lillian Lee Department of Computer Science Cornell University Ithaca, NY 14853 USA, Shivakumar Vaithyanathan IBM Almaden Research Center 650 Harry Rd. San Jose, CA 95120 USA," Thumbss up? Sentiment Classification using Machine Learning Techniques " ProceedingEMNLP '02 Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10Pages 79-86-2002.

[14] Andranik Tumasjan, Timm O.Sprenger,Philipp G.Sandner, Isabell M.Welpe, Technisehe Universitat Munehen, Lehrstuhl fur Betriebsehaftslehre 139,80804 Munich, Germany," Predicting Election With Twitter:what 140 Characters Reveal about Political Sentiment" Proceedings of the Fourth International AAAI Conference on Webblogs and social Media-2010.

[15] Francis P. Barclay, Pichandy Chinnasamy, and Priyadarshni Pichandy," Political Opinion Expressed in Social Media and Election Outcomes - US Presidential E" GSTF International Journal on Media & Communications(JMC) Vol.1 No.2, February 2014.

[16] Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar and Shrikanth Narayanan," A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle." Annenberg Innovation Laboratory (AIL), Signal Analysis and Interpretation Laboratory (SAIL), University of Southern California, Los Angeles, CA. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 115–120, Jeju, Republic of Korea, 8-14 July 2012. c 2012 Association for Computational Linguistics.

[17] Lei Shi, Neeraj Agarwal, Ankur Agrawal, Rahul Garg, Jacob Spoelstra"Predicting US Primary Elections with Twitter",2012

[18] Ravi Parikh,Mattine Movassate," Sentiment Analysis of User-Generated Twitter Updates using Various Classification Techniques",June 4,2009.

[19] Marko Skoric, Nathaniel Poor, Palakorn Achananuparp, Ee-Peng lim,Jing Jiang," Tweets and Votes:A study of the 2011 Singapore General Election",45th Hawaii International Conference on System Sciences-2012.

.http://www.analyticsvidhya.com/blog/2014/10/introduction-neural-network-simplified

