International Journal of Innovative Research in Science and Engineering
Vol. No.2, Issue 09, September 2016
www.ijirse.com
ISSN (O) 2454-9665
ISSN (P) 2454-0663

AN EFFICIENT CLUSTERING TECHNIQUE FOR CLASSIFICATION OF TWITTER DATA FOR DATA ACCURACY USING SENTIMENTAL ANALYSIS

Vishal C[1], Dr. K. Saravanan[2], B.H.Chandra Shekar[3]

[1]*Research Scholar, PRIST University, Thanjavur, Tamilnadu, (India)*

[2]*Dean - Faculty of Computer Science,PRIST University, Thanjavur, Tamilnadu, (India)*

[3]*Associate Professor, Dept of MCA, RV College of Engineering, bangalore, (India)*

ABSTRACT

The explosive growth of social media on the Web, individuals and organizations are increasingly using the content in these media for their decision making. Twitter is a highly popular social networking and micro-blogging service used by millions worldwide. Each status or tweet is a maximum of 140-160 character text messages. Registered users can read and post tweets, but unregistered users can only view them. Human beings express their feelings with the help of words, in the form of speech or comments on any social media. Analyzing customer reviews and Sentiment is most important, we Classify the customer's reviews into three classes: Neutral, Positive and Negative Sentiment. An efficient clustering technique is proposed by using Wards method to get accurate dendrograms. This method removes duplicate attributes and lead to best classification for data accuracy.

Keyword: Big Data, Sentiment Analysis, Twitter

I. INTRODUCTION

The increasing popularity of social media (such as online communities) has spawned huge amount of information in the society today. Analytics of social media data helps to maximize its utility, which plays a significant role in the exploration of data. Sentiment analysis plays a vital role in the analysis of the real time using natural language processing technique. This technique can be applied for analyzing real time data generated by various social Medias. Twitter is a highly popular social networking and micro-blogging service used by millions worldwide. Each status or tweet is a maximum of 140-160 character text messages.

Registered users can read and post tweets, but unregistered users can only view them. Human beings express their feelings with the help of words, in the form of speech or comments on any social media. To understand what people need or think about the particular object is very important in this society. Analyzing customer review is most important, by doing that we tend to rate the Product and provide opinions for it which is been a challenging problem today.

.

ISSN (O) 2454-9665
ISSN (P) 2454-0663

II. BACKGROUND WORK

In [1], Nasukawa and Yi. illustrate a sentiment analysis approach to extract sentiments associated with polarities of positive or negative for specific subjects from a document, instead`112345R of classifying the whole document into positive or negative. The essential issues in sentiment analysis are to identify how sentiments are expressed in texts and whether the expressions indicate positive (favorable) or negative (unfavorable) opinions toward the subject. Powerful functionality for these kinds of issues is used.

In [8], Ding et al. proposed an effective method for identifying semantic orientations of opinions expressed by reviewers on product features. It is able to deal with two major problems with the existing methods, (1) opinion words whose semantic orientations are context dependent, and (2) aggregating multiple opinion words in the same sentence. For (1), a holistic approach is proposed that can accurately infer the semantic orientation of an opinion word based on the review context. For (2), a new function to combine multiple opinion words in the same sentence is proposed.

Taylor et al. [9] presented a generic design of a tourism opinion mining system that aims to be useful in many industries. They also used their proposals to successfully implement the system and solve a specific problem in the Lake District tourism industry.

In Zhu et al [10] an aspect-based opinion polling system takes as input a set of textual reviews and some predefined aspects, and identifies the polarity of each aspect from each review to produce an opinion poll.

In [11] Haddi, Lui and Shi investigated the sentiment of online movie reviews. They used a combination of different pre-processing methods to reduce the noise in the text in addition to using chi-squared method to remove irrelevant features that do not affect its orientation. Authors have reported extensive experimental results, showing that, appropriate text pre-processing accuracy achieved on the two data sets is comparable to the sort of accuracy that can be achieved in topic categorization, a much easier problem.

In Moraes, Valiati & Neto [12] they have focused on comparing SVM and ANN in terms of the requirements to achieve better classification accuracies. In this, experiments evaluated both methods as a function of selected terms in a bag-of-words (unigrams) approach. Regarding the sentiment learning literature, the main findings/contributions are in the two points. First point is in terms of classification accuracy on the benchmark dataset of Movies reviews and second point as an overall comparison in the context of balanced data.

III. DIFFERENT LEVELS OF SENTIMENT ANALYSIS

3.1 Document level

The task at this level is to classify whether a whole opinion document expresses a positive or negative sentiment For example, given a product review, the system determines whether the review expresses an overall positive or negative opinion about the product. This task is commonly known as document-level sentiment classification.

3.2 Sentence level

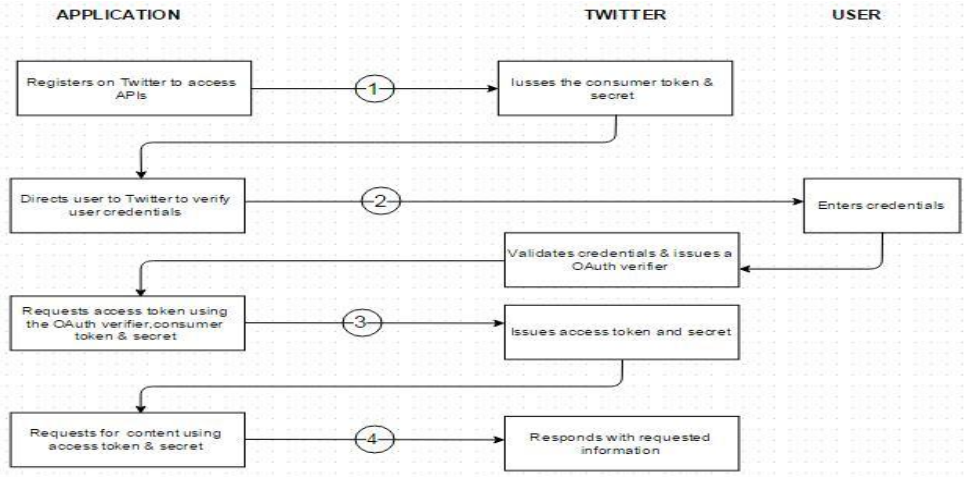
The task at this level goes to the sentences and determines whether each sentence expressed a positive, negative, or neutral opinion. Neutral usually means no opinion. This level of analysis is closely related to subjectivity classification which distinguishes sentences (called objective sentences) that express factual information from sentences (called subjective sentences) that express subjective views and opinions.

International Journal of Innovative Research in Science and Engineering
Vol. No.2, Issue 09, September 2016
www.ijirse.com
ISSN (O) 2454-9665
ISSN (P) 2454-0663

3.3 Entity and Aspect level

Both the document level and the sentence level analyses do not discover what exactly people liked and did not like. Aspect level performs finer-grained analysis. Aspect level was earlier

Called feature level (feature-based opinion mining and summarization), aspect level directly looks at the opinion itself. It is based on the idea that an opinion consists of a sentiment (positive or negative) and a target (of opinion).



[1]The Architecture diagram of OAuth

Applications are also known as consumers and all applications are required to register themselves with Twitter. Through this process the application is issued a consumer key and secret which the application must use to authenticate itself to Twitter. The application uses the consumer key and secret to create a unique Twitter link to which a user is directed for authentication. The user authorizes the application by authenticating himself to Twitter. Twitter verifies the user's identity and issues a OAuth verifier also called a PIN. The user provides this PIN to the application. The application uses the PIN to request an "Access Token" and "Access Secret" unique to the user. Using the "Access Token" and "Access Secret", the application authenticates the user on Twitter and issues API calls on behalf of the user.

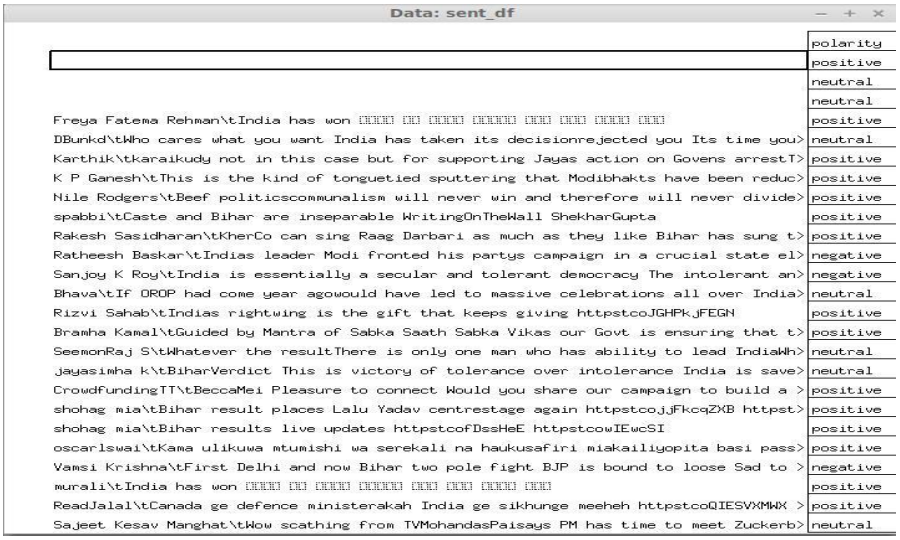
IV. DIVISIVE METHOD

In this method we assign all of the observations to a single cluster and then partition the cluster to two least similar clusters. Finally, we proceed recursively on each cluster until there is one cluster for each observation.

4.1 Dendograms

The dendrograms are created to visualize the result of a hierarchical clustering calculation. To Count the frequency of the word and create from the tweets after cleaning the data. It also counts the word by plotting word frequency. Clustering algorithm is used for dendrograms by using Ward's method.

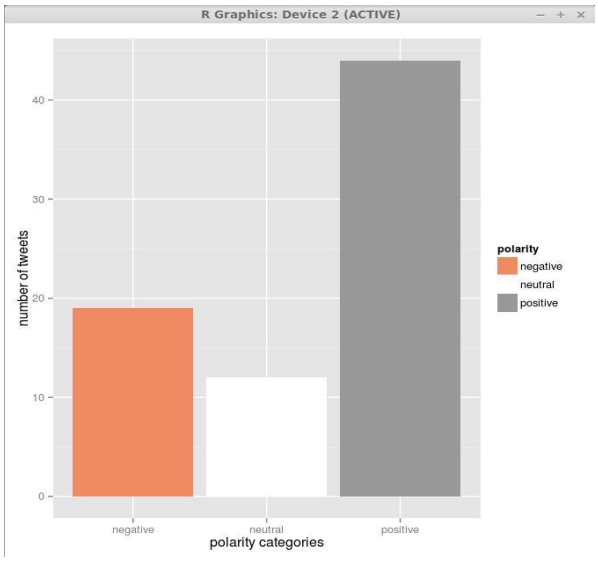
International Journal of Innovative Research in Science and Engineering
Vol. No.2, Issue 09, September 2016
www.ijirse.com
ISSN (O) 2454-9665
ISSN (P) 2454-0663



[2]Figure shows the data analysis page which contains the tweets classified into different polarity class.

V. RESULTS AND DISCUSSIONS

The twitter sentiment analysis gives the result based on the tweets, retweets for a particular keyword. The positive and negative words are matched with data collected after mining [**12]**. The process of sentiment analysis contains two text files named positive.txt and negative.txt. The positive file and contains the general positive words in English like accurate, better, calm, decent, glad, honest etc. The negative file contains the general negative words in English like abort, danger, pain, refuse, sad, waste etc. name, time, date, score, results. Later the tables are connected to analyse the data.



[3] Figure of statistics of tweets classified into three categories namely positive, negative and neutral in the form of bar chart

ISSN (O) 2454-9665
ISSN (P) 2454-0663

Data dictionary with two separate files containing positive and negative words. The sentiment analysis applies to the tweets. The positive file and contains the general positive words in English like accurate, better, calm, decent, glad, honest etc. The negative file contains the general negative words in English like abort, danger, pain, refuse, sad, waste etc [18]. Arrange them in alphabetical order. Count the number of positive and negative words find differences between them check whether it is >0 or <0. If the result is positive, the people are talking, tweeting good things on that topic. If the result is negative, the people are talking, tweeting bad things on that topic

VI. CONCLUSION

Sentiment analysis system hence analyses the data collected from twitter on word basis and classifies into positive, negative and neutral. The number of positive and negative words in the data dictionary may increase, which are present in the files with the accurate meaning. This classification Analyzes customer review of a product and helps in decision making.

REFERENCES

[1] T. Nasukawa, "Sentiment Analysis: Capturing Favorability Using Natural Language Processing Definition of Sentiment Expressions," pp. 70–77, 2003. 3. Mikalai Tsytsarau, Themis Palpanas "Survey on mining subjective data on the web" Data

[2] Min Knowl Discov, 24 (2012), pp. 478–514

[3] Efthymios Kouloumpis,Theresa Wilson and Johanna Moore,2012: " Twitter Sentiment Analysis: The Good the Bad and the OMG!." Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.

[4] Bifet, A., and Frank, E. 2010. "Sentiment knowledge discovery in twitter streaming data", Proceedings of 13th International Conference on Discovery Science.

[5] Zhu Wei-ping, Li Ming-xin, Chen Huan, "Using MongoDB to Implement Textbook Management System instead of MySQL",IEEE,978-1-61284-486,2011.

[6] Apoorv Agarwal, Boyi Xie and Rebecca Passonneau(2012). "Sentiment Analysis of Twitter Data". In Proceeding to LSM "11 of the workshop on Languages in Social Media Pages 30-80.

[7] X. Ding, S. M. Street, B. Liu, S. M. Street, P. S. Yu, and S. M. Street, "A Holistic Lexicon-Based Approach to Opinion Mining," pp. 231–239, 2008.

[8] E. Marrese-Taylor, J. D. Velasquez, and F. Bravo-Marquez, "Opinion Zoom: A Modular Tool to Explore Tourism Opinions on the Web," 2013 IEEE/WIC/ACM Int. Jt. Conf. Web Intell. Agent Technol., pp. 261–264, Nov. 2013.

[9] J. Zhu, H. Wang, M. Zhu, B. Tsou and Matthew M, "Aspect-Based Opinion Polling from Customer Reviews", "IEEE Transaction On Affective Computing", vol. 2, NO. 1, January-March 2011.

[10] E. Haddi, X. Liu, and Y. Shi, "The Role of Text Pre-processing in Sentiment Analysis," Procedia Comput. Sci., vol. 17, pp. 26–32, Jan. 2013.

[11] R. Moraes, J. F. Valiati, and W. P. Gavião Neto, "Document-level sentiment classification: An empirical comparison between SVM and ANN," Expert Syst. Appl., vol. 40, no. 2, pp. 621–633, Feb. 2013.

ISSN (O) 2454-9665
ISSN (P) 2454-0663

[12] E. Haddi, X. Liu, and Y. Shi, "The Role of Text Pre-processing in Sentiment Analysis," Procedia Comput. Sci.,

[13] vol. 17, pp. 26–32, Jan. 2013.

[14] R. Moraes, J. F. Valiati, and W. P. Gavião Neto, "Document-level sentiment classification: An empirical comparison between SVM and ANN," Expert Syst. Appl., vol. 40, no. 2, pp. 621–633, Feb. 2013.

Biography of authors

Dr. K. Saravanan received his M.Sc. in Computer Science from A. V. C. College (Autonomous), Mayiladuthurai in 1992, M.S. in Software Systems from B.I.T.S. Pilani in 1998, M.Phil. in Computer Science from M.S. University, Tirunelveli in 2003, Ph.D. in Computer Science from PRIST University, Thanjavur in 2011 and M. Tech., in CSE from PRIST University, Thanjavur in 2013. He is having 24+ years of teaching experience and 13+ years of research experience. He guided more than 70 candidates at M.Phil level and guiding 8 candidates at Ph.D. level. Right now, He is working as Dean, Faculty of Computer Science, PRIST University, Thanjavur. His research interest includes in the areas of Big Data Analytics, Data Mining, Cloud Computing, Wireless Sensor Networking and Computational algorithms.

Mr.vishal C received M.C.A. Degree in Computer Application from R.N.S.I.T Bangalore in the year 2006, perusing (Ph.D.) in Computer Applications from PRIST University, Thanjavur. Having 8+ years of teaching experience and 2 years of industry experience. Working as Assistant Professor, Faculty of Master of Computer Applications, RV College of Engineering, Bangalore. My area of research interest includes Big Data Analytics, Data Mining, Cloud Computing.