© June 2016 | IJIRT | Volume 3 Issue 1 | ISSN: 2349-6002

Analyzing Twitter data for Multiple languages by applying machine learning Techniques

Manu Krishna Bhardwaj, Brajesh Kumar
Computer Science and Engineering,
Manav Rachna College of Engineering, Faridabad, India

***Abstract*— Most of the people are using social network sites these days to express their ideas/views about any event or product. These ideas are in English language. But in a Country like ours these may be in multi lingual because there are fifteen languages recognized in our constitution. In this paper we implement a text mining process on the basis of multi-disciplinary language by apply machine learning tool and techniques. So the performance of opinion mining increases by using Multiple Languages as compare to single language.**

***Index Terms*— Positive Word, Negative Word, Word Cloud, Bayes theorem, Classification, Hindi Words, Hindi Spelled in English Words**

I. INTRODUCTION[8]

Micro blogging sites are social media sites which act as social network for users to communicate short and limited character posts. The highest credible micro blogging social networking site is twitter on which user can read and post tweets which are 140-150 characters limit in length. The twitter posts and comments normally called as tweets. Industries use sentiment analysis to know the opinion of customers about their product, this helps to raise the productivity and marketing of the company.

The twitter data known as twitter corpus gives free knowledge about any event in the form of stream. This Knowledge has leads to a multiplicity of new work, eg. Opinion of new released movie, customer views on update or new product, analysis of daily monthly and yearly share market, estimation of public opinion during elections session. These sentiment analysis attempts important role play for online survey of any product or event. Analysis gives fruitful result for the business performance growth, monitoring from customer needs and ideas other than making door to door surveys, which are expensive and time consuming [1].

Hence for automated classification of opinions, analysis tool using machine learning techniques into subjective (positive or negative) aspects of public, the twitter data is most useful.

II. DATA SOURCE OF OPINION MINING [8][6]

The customers and industries of various fields use the same source of data for various purposes and exploit it. The main decisive factor for the enrichment of the quality services given and improvement of deliverables are the opinion of user customers. There are three levels of good understanding about products and events are Review sites, blogs and micro blogs.

A. Review Sites[6]

It is important for the users to take the opinion before making a purchase because opinion is the actual data for taking decision in making purchase. The views in respect of the products or services are generally available on web. The classification of opinion of the users/reviewers is obtained and fetched from the web. E.g. Shopping products reviews on www.flipkart.com, www.snapdeal.com etc.

B. Blogs [6]

The blogosphere is the name of all the blogs sites and services attached to the word. Blogs are the explanation of views of the individual users at multiple events. These are as like diary of user.it is free and un-ended in nature. It has no character limitations. Blogs are the best source of expression of views and ideas during the studies which are related to the sentiment analysis. [6][7].

C. Micr-Blogging [6]

Micro logs are well known equipment of communication for social networking website users. The comments and reviews on Facebook or Twitter pages form as a best data source. The micro- blog sites have some character limit. The large number of small comments used by users or customs is in the twitter. In micro blog sites the multiple users can gives their views and ideas on an event. Innumerable messages come every day on this popular website for micro blogging which are Facebook, Google+ and Twitter

III. MOTIVATION[6]

Customer ideas, reviews and comments on an event or product may be varied language i.e. (English, Hindi, Local Language spelled in English etc.), which is hard task to understand the

sentiment in varied language. There is a queer mixture of noun and verbs from which the meaning cannot be constructed properly example in Hindi words spelled in English are like dena bank m cheque mat dena.so 'dena' has been used twice as verb as well as noun word.

The method of using Lexicon based analysis for opinion mining is not an impressive in use of context dependent words. E.g. 'small' word consider as positive and negative sentiment of a company product feature. For a mobile phone if customer review that ―size of mobile phone is small‖ this sentence does not show either size is positively opinioned or negatively[6]

IV. OPININON ANALYSIS

Sentiment analysis was first introduced by Liu, B [5].The products or services are valued by people through their expressions of opinions, ideas, on the twitter this known as subjective analysis. Subjective means the opinion of users consider in some negative, positive and neutral aspects. If it is neutral the words not consider in classification. the neutral or use less words are neglected. The words which have no meaning or not consider during sentiment classification are objective analysis. Sentiment analysis may be used in any three levels of sites. Micro-blogs messages are limited character limit up to 149 characters per micro blogs. The analysis done in three levels word level sentiment analysis, sentence level sentiment analysis and document level sentiment analysis.[12].Word level used for producing word based opinion summery of multiple or varied reviews. Sentence level analysis determined analyses of whole sentences consider as positive negative or neutral opinion. Sentence gives no opinion means it is an objective analysis which are neglected during classification. This level of analysis is closely related to subjectivity classification (Wiebe, Bruce and O'Hara, 1999), which distinguishes sentences (called objective sentences) that express factual information from sentences (called subjective sentences) that express subjective views and opinions [12]. This level of analysis assumes that each document expresses opinions on a single entity [12].

V. RELATED WORK

Due to its widespread and popularity, Eman says that there is a innumerable user reviews or comments on a product produced and shared every day. In his work, an open source perspective is submitted, throughout which, twitter tweets data has been fetched ,these tweets are called micro –blogs. Then worked on this collected micro bogs and pre-processed this micro blogs, after preprocessing he analyzed the score of sentiment and analyzed score visualized by graph of using open source tools to perform text mining and sentiment analysis for analyzing a case study on reviews of about two giant retail stores in the UK namely Tesco and Asda stores over Christmas period 2014[1]. Collecting customer reviews door to door can be expensive and time consuming task; such as surveys. This method is known as conventional methods. The sentiment analysis of the customer opinions makes it easier for businesses perspectives for analysis competition of market products to compare the product value with their competitor, which also provide an insight into future marketing strategies and decision making policies [1].

AmrutaTarlekar and Kodmelwar M.K studied about advantageous to researchers in political data for opinion mining domain, learning technologies and learning analytics. It provides a workflow for analyzing social issues, social media data on political domain which overcomes the major drawback of both manual qualitative analysis and large scale computational analysis of user generated textual content. The system can be extended to do sentiment analysis on any domain like movie, songs, and product feedbacks just by changing training corpus [2].

Alexandra Balahur, Marco Turchi,they worked on an challenging task which was sentiment analysis from twitter micro-blogs in a multi-language dictionary words setting. They built sentiment analysis system for tweets using English word nets only. After that, the data translated from English to four other languages - Italian, Spanish, French and German - using a machine translation. Further on, we manually correct the test data and create Gold Standards for each of the target languages. Finally, they check the performance of the analysis data by classifiers for the different languages concerned and show that the joint use of training data from multiple languages (especially those pertaining to the same family of languages) significantly raise the analysis results of the sentiment classification[23].

In many cases multilingual text for analysis is required and a simple translation of the text to English would result in worthless solutions. A novel field of application is the analysis of the communication in social media by politicians in a country with multiple national languages, such as Switzerland [24].

Lucas Brönnimann, did the work by applying machine learning techniques using case study on Swiss politicians review. This approach was applied to determine the connection with a party for anyone writing about political subjects on Twitter[24]. While text similarity alone achieves acceptable results, it can be shown that the combination with a multilingual sentiment analysis for the key topics raise the performance of system as compare to single language sentiment approach. He also derived an sentiment algorithm in

which emoticon symbols used to determine the given text is positive or negative This allows for a language specific acquisition of a sentiment lexicon which can be used with a simple algorithm to determine the sentiment of Messages on Twitter in their respective language [24].

VI. RESOURCES

Hindi, English and Hindi words Spelled in English language are rear to find in combined for use as Dataset Recourses. And it is difficult to work on multi lingual dataset. Our aim is to build dataset resources of multiple languages (Hindi, English and Hindi words Spelled in English language) and improve the performance of sentiment analysis of any event ideas or reviews of people. The resultant values of multiple resources are much better than using single Language Dataset resources. The performance of Analytical system will be improved as compared to the single language data set resources.

A. Data Set From Twitter

We retrieved 1500 micro blogs dataset from tweeter on a case study about any event. Each tweet contains 140 to 150 characters limit which are separated in different class levels (positive, negative and neutral) We designed word cloud of these micro blogs words which given in Fig 1.The cloud contain both subjective data set and neutral data set. Subjective data set have some meaning or build a result in two ways either resultant score is negative or positive while neutral data is being neglected from the comment and not participate in clarification of the result.

Fig.1. ***Micro-blog clouds***

B. Dictionary of Combined Word Net

In this context the subjective data is a data which contain negative or positive multilingual words. In a day or two continuously it is possible to collect sufficient both negative and positive data but the subjective data was collected on different days to check the performance of our case study. If the data are collected on same time, it is considered good because it can capture substantial difference between positivity and negativity of any activity. The multilingual used to collect the data is shown in:

Language	POSITIVE	NEGATIVE
Hindi Word Net	सर्वश्रेष्ठ, सुधार, नियंत्रित, सर्वाधिक	नहीं, अंतिम, परेशान, पुराना
Hindi words spelled in English word Net	bahaadur, saavadhan, satark, chaalak	Naaraaj,aakraamak, kashtdaayi ,abhimaani
English Word Net	Abound, calm, brotherly, idyllic	Ache, babble, crook, giddy

Table1. Few data sets dictionary words of multilingual words.

The data sets of negative and positive words were used simultaneously. This shows the Micro blogs of twitter matches with the negative and positive word net or both then this will be a subjective data. In our Case Study work 1500 subjective Micro blogs out of which 505 of negative and 306 negative and 689 are those which contains both negative and positive in equal form by using multilingual word net.

VII. CLASSIFICATION ALGORITHM [25]

In our work of analysis we classify our analysis scoring using Naive Bayesian classification. It is simple and probabilistic classification initially based on Bayes theorem the following Bayes theorem.

Let X be the complete Set and A and B the set of Positive and Negative Reviews respectively then

$$P(A) = \frac{\text{Number of favorable outcomes to A}}{\text{Total number of outcomes}} \quad \#\text{Probability of A}$$

$$P(B) = \frac{\text{Number of favorable outcomes to B}}{\text{Total number of outcomes}} \quad \#\text{Probability of B}$$

the Probability of B given that Event A has already happened.

$$P(A|B) = \frac{P(A)\,P(B|A)}{P(B)}$$

the Probability of A given that Event B has already happened.

$$P(B|A) = \frac{P(B)\,P(A|B)}{P(A)}$$

The naïve Bayes algorithm uses Bayes theorem. The formula P (A|B) states the conditional probability of A given B, where A is a class label and B a feature. It allows calculating unknown conditional probability form a known conditional probability together with the prior probabilities. It is assumed that the presence of a feature is unrelated to the presence of any other feature .[25]

VIII. PROPOSED METHODOLOGY

We are using R programing tool i.e Rstudio Version 0.99.879 - © 2009-2016 RStudio, Inc.The steps involved in the methodology are described as follows:

A. Extract data from Twitter for collecting Data Set:

I. Create application on tweeter in this link below:
https:// apps.teitter.com/app/new

II. Generate Consumer keys of this Application.
Consumer Key: The consumer key provided by your application.
Consumer Secret: The consumer secret provided by your application.

III. Generate Access Keys of Same Application
Request URL:The URL provided for retrieving request tokens
Auth URL: The URL provided for Authorization/verification purposes.
Access URL: The URL provided for retrieving access tokens.
Oauth Key: For internal use by machine for generating keys for access rights.
Oauth Secret: For internal use by machine for generating keys for access rights securely.

B. Collect and Build Data Set from Twitter Tweets
This collecting data set called Micro Blogs. In our work we are collecting #ODDEVEN Scheme in Delhi word Data set. We collect 1500 micro blogs (i.e n=1500) by. searchTwitter('#oddeven', n=1500).

C. Scan Dictionary files: Dictionary contains all Hindi English and Hindi and Hindi spell in English words.

D. Find frequent words and associations.

E. Comparing Score of single language and multi languages

The score of single language analysis word net with Micro Blogs and generates a normalized table of negative positive and zero (neutral) scores. Using single language dictionary for analysis ODDEVEN the neutral data score comes 943.So out of 1500 micro blogs which we build using twitter tweets we have got 943 of data which have no meaning or the words are neglected during analysis. So it concludes that if we use only single lingual dataset the performance of score is low. The compare data show in table below:

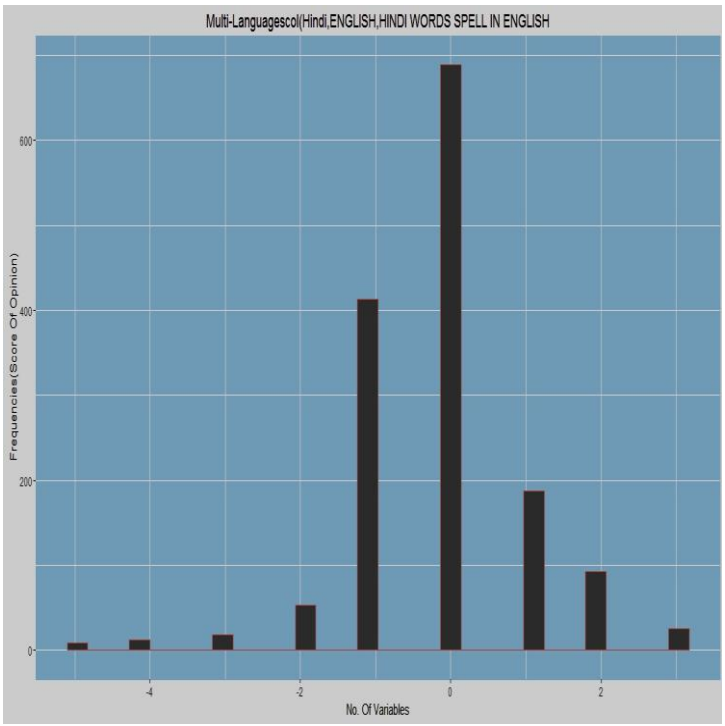

*Fig. 2.**Using Single Language data Dictionary of data Score to show the Grammar of Graphical form..***

Table2. Matrix data combine the comparison between single language analysis data and Multilingual Analysis Data.

Data set Score Language	Negative					Neutral	Positive		
	-5	-4	-3	-2	-1	0	1	2	3
English Language	5	0	1	30	296	943	143	30	52
English + Hindi + Hindi words Spelled in English	9	12	18	53	413	689	187	93	26

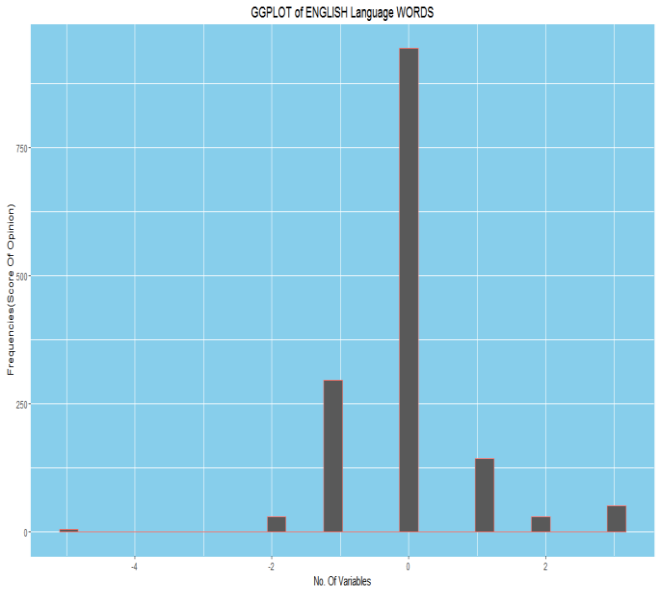

Fig.3. ***Using Multiple Language data Dictionary of data Score to show the Grammar of Graphical form.***

F. Plot Show Results of Normalized ODD EVEN data as shown in figure

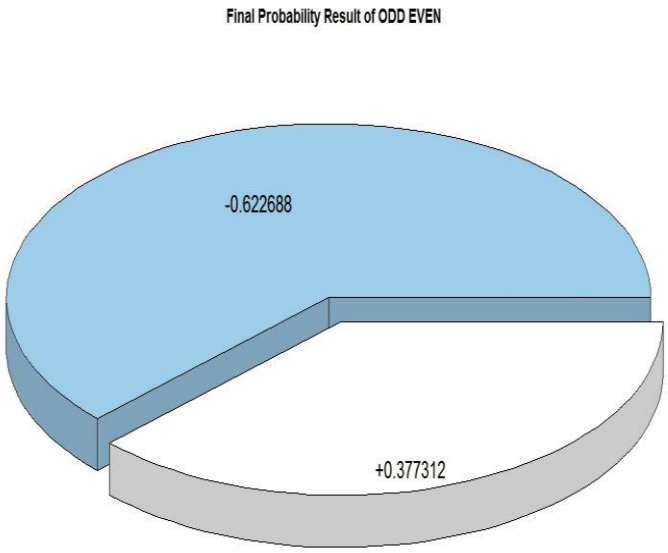

Fig.4. ***Using Multiple Language data Dictionary of data Score to show the Classification by using Naïve Bayesian classification Technique .***

When we using single language data source the graph of analysis score of odd even scheme data shows 943frequencies or results of micro blogs goes to neutral means no use during classification subjective score is low shown in fig. 2.on the other hand if we use multilingual data the performance is raise and the subjective data values increases and neutral analysis data is 689 out of 1500 micro blogs which is less than single lingual data analysis so by using multilingual data the performance of analysis system raise in Fig. 3 .

G. Classification on Basis of Algorithm

We are classifying the score using Naïve Bayesian Algorithm. Even Odd data set of our given case study is being classified by Naïve Bayesian classification algorithm and gives the best performance results of both negative and positive form.

From the perusal of this pie chart in Fig.4. the final classification result of negativity (0.622688) is greater than the positivity (0.377312). So the result of twitter survey in respect of ODD_EVEN scheme comes negative according to my classification work.

IX. CONCLUSION AND FUTURE WORK

The most of the comments on social media sites would be written in different languages for the ease of public .in this proposed research is to analysis twitter data with hindi words, hindi words spelled in english and english language. After

evaluating the tweets through naïve bayesian we compared with survey data news 18 which is almost similar with my result. The sentiment analysis has been done on the feedback of any activity in some multidisciplinary languages like hindi, english from different social networking sites gives better performance as compared to english language in india.

REFERENCES

[1] Eman M.G. Younis, " Sentiment Analysis And Text Mining For Social Media Microblogs Using Open Source Tools: An Empirical Study", International Journal Of Computer Applications(0975 – 8887), February 2015, Volume 112 – No. 5.

[2] AmrutaTarlekar, Prof. Kodmelwar M.K, "Sentiment Analysis of Twitter Data from Political Domain Using Machine Learning Techniques", International Journal of Innovative Research in Computer and Communication Engineering.,ISO 3297: 2007 Certified Organization, June 2015,Vol. 3, Issue 6.

[3] S. Kim,E. Hovy, "Determining the sentiment of opinions", In Proc. 20th Int. Conf. Comput.l Linguistics, PA, USA, 2004, pp. 1367–1363.

[4] Lina L. Dhande ,Dr. Prof. Girish K. Patnaik, "Analyzing Sentiment of Movie Review Data using Naive Bayes Neural Classifier", International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 3, Issue 4, July-August 2014, ISSN 2278-6856.

[5] Liu, B. (2010). Sentiment analysis and subjectivity. Handbook of natural language processing, 2, 627-666.

[6] Nidhi R. Sharma ,Vidya D. Chitre, "Opinion Mining, Analysis and its Challenges", International Journal of Innovations & Advancement in Computer Science,ISSN 2347 – 8616,Volume 3, Issue 1,April 2014.

[7] Zhai Z, Liu B, Xu H, and Jia P, "Grouping Product Features Using Semi-supervised Learning with Soft-Constraints", In Proceedings of COLING. 2010.

[8] Manu Krishna Bhardwaj, Brajesh Kumar, "Opinion Mining Of Social Media Data Using Machine Learning Techniques‖", International Journal of Scientific Engineering and Applied Science (IJSEAS) ,Volume-2, Issue-5, May 2016, ISSN: 2395-3470.

[9] ZhongchaoFei, Jian Liu, and Gengfeng Wu: "Sentiment Classification Using Phrase Patterns", Proceedings of the Fourth International Conference on Computer and Information Technology in 2004.

[10] Wilson, T., Wiebe, J. and Hwa, R, "Just how mad are you? Finding strong and weak opinion clauses", Proceeding of National Conference on Artificial Intelligence in 2004.

[11] Hiroshi, K., Tetsuya, N., and Hideo, W. "Deeper sentiment analysis using machine translation technology". In Proceedings of the 20th international Conference on Computational Linguistics (Geneva, Switzerland) in 2004.

[12] B. Liu, "Sentiment Analysis and Opinion Mining", Morgan & Claypool Publishers, April 22, 2012.

[13] Patil Monali S1, Kankal Sandip, "A Concise Survey on Text Data Mining", International Journal of Advanced Research in Computer and Communication Engineering,Vol. 3, Issue 9, September 2014.

[14] Vandana Korde and C. Namrata Mahender, "Text Classification and Classifiers :A Survey", International Journal of Artificial Intelligence & Application, Vol.3, No.2, March 2012.

[15] N. Kanya and S. Geetha, "Information Extraction: A Text Mining Approach",International Conference on Information and Communication Technology in Electrical Sciences, IEEE (2007).

[16] Guoliang Li, Beng Chin Ooi, Jianhua Feng ,Jianyong Wang and Lizhu Zhou, " EASE: An Effective 3-in-1 Keyword Search Method for Unstructured, Semi-structured and Structured Data".

[17] Charles L. Wayne, "Topic Detection & Tracking (TDT) Overview & Perspective".

[18] G.Salton, A. Wong, and C. S. Yang." A vector space model for automatic Indexing",Communications of the ACM, 18(11):613–620, 1975.

[19] Sushmita Mitra, Tinku Acharya " Data Mining Multimedia, Soft Computing, and Bioinformatics".

[20] Jonathan G. Fiscus and George R. Doddington, " Topic Detection and Tracking Evaluation Overview".[18] Charu C Aggrawal and Chengxiang Zhai,"Mining Text Data".

[21] David M Blei, Princeton University, "Introduction to Probabilistic Topic Models".

[22] Vishal Gupta and Gurpreet S. Lehal, "A Survey of Text Mining Techniques and Applications", Journal of Emerging Technologies in Web Intelligence, Vol. 1, No. 1, August 2009.

[23] Alexandra Balahur, Marco Turchi, "Improving Sentiment Analysis in Twitter Using Multilingual Machine Translated Data", Proceedings of Recent Advances in Natural Language Processing, pages 49–55,Hissar, Bulgaria, 7-13 September 2013.

[24] Lucas Brönnimann, ―Multilanguage sentiment-analysis of Twitter data on the example of Swiss politiciansǁ, University of Applied Science Northwestern Switzerland, CH-5210 Windisch, Switzerland

[25] Reshma Bhonde, Binita Bhagwat , Sayali Ingulkar and Apeksha Pande, "Sentiment Analysis Based on Dictionary Approach", International Journal of Emerging Engineering Research and Technology ,Volume 3, Issue 1, January 2015, PP 51-55 ISSN 2349-4395 (Print) & ISSN 2349-4409