International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 2, Issue 1, ISSN No. 2455-2143, Pages 67-72
Published Online November-December 2016 in IJEAST (http://www.ijeast.com)

REVIEW PAPER ON SENTIMENT CLASSIFICATION OF MOVIES REVIEW.

Shivangi Sharma
Research Scholor (CSE)
S.S.C.E.T Badhani
Pathankot, India

Gurjeet Kaur
Assistant Professor (CSE)
S.S.C.E.T Badhani
Pathankot, India

.

***ABSTRACT* - The process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral. The papers are basically focussed on encapsulating the movie reviews at characteristic level so that user can find easily that which character of the movie they liked or disliked. The system also performs subjectivity analysis. The subjective analysis is one and most useful tasks that take place in sentiment analysis. Online reviews have mostly consists of objective and subjective sentences. Talking about the objective sentences, it mostly consists of factual information and no opinion or sentiments. On the other hand, subjective sentences broadly cover the greater interpretation based on personal feeling, emotions, aesthetics etc to find summary at the characteristic level. In this paper, two different methods are implemented for finding subjectivity of sentences and then rule based system is used to find feature-opinion pair and finally the orientation of extracted opinion is revealed using two different method. Initially the proposed system uses SentiWordNet approach to find out orientation of extracted opinion and then it uses the method which is based on lexicon consisting list of positive and negative words.**

Keywords**:-data mining, Sentiment analysis, subjectivity, objectivity, sentiword.net, Naïve bayes**

I. INTRODUCTION

Nowadays, Social media is becoming more and more popular since mobile devices can access social network easily from anywhere. Therefore, Social media is becoming an important topic for research in many fields. As number of people using social network are growing day by day, to communicate with their peers so that they can share their personal feeling every day and views are created on large scale. Social Media Monitoring or tracking is most important topic in today's current scenario. In today many companies have been using Social Media Marketing to advertise their products or brands, so it becomes essential for them that they can be able to calculate the success and usefulness of each product .[2]For Constructing a Social Media Monitoring, various tool has been required which involves two components: one to evaluate how many user of their brand are attracted due to their promotion and second to find out what people thinks about the particular brand Humors, that have been generated can be evaluated usually by performing various Key performance factors such as the number of followers or friends, the number of likes or shares or comment for each post and more difficult one like engagement rate, response time to evaluate them and other composite measures. Measuring the Large dataset is usually direct and can be done by using some statistical method.On the other hand, to evaluate the opinion of the users is not as easy as it seems to all users. For evaluating their attitude may requires to perform Sentiment Analysis, which is defined as to identify the polarity of customer behavior, the subjective and the emotions of particular document or sentence.

To process this we need Machine Learning and Natural Language Processing methods and this is place where most of the developers facing difficulty when they are trying to form their own tools.Over the recent years, an emerging interest has been occurred in supporting social media analysis for advertising, opinionanalysis and understanding community cohesion. Socialmedia data adapts to many of the classifications attributed for "big-data" – i.e. volume, velocity and variety.

Analysis of Social media needs to be undertaken over large volumes of datain an efficient and timely manner.Analysing the media content has been centralized in social sciences, due to the key role that the social media plays in modelling public opinion. This type of analysis typically on thepreliminary coding of the text being examined, a step that involves to read and annotate the text, and that limits the sizes of the data that can be analysed.

II. LITERATURE SURVEY

1) Ana Mihanovic, HrvojeGabelica, ZivkoKrstic. "***Big Data and Sentiment Analysis using Knime: Online Reviews versus Social media***", (**2014)-**This paper analyze sentiment analysis on various gadgets in two different forms i.e. online review and Tweets. For dictionary making, it uses Knime tool in both forms. Online reviews have been crawled using Apache Nutch crawler while tweets were collected using Java package. As tweets are shorter so, number of tweets collection will be more compare to online reviews. Both tweets and online review are stored in HBase table on Apache Hadoop server. Data sets for online review are classified based on key, PID, Review Date, Review Text, Keyword, Language while Tweets are having attribute such as Key, UserScreenName, Creation Date, Text, Keyword, Language. This data is loaded into Knime. Dictionary build for online review can be easily categorized based on usage, price, quality, experience of user, Look and services provided by gadgets. Correct grading of phrases are to be done using grade scope. For tweets, Scoring is done based on polarity i.e. positive, negative or neutral. For tweets it will be difficult for scoring based on phrases as it is impossible to categorize them.

They need more preprocessing and it uses frequency driven. So this paper concludes that Sentiment analyses on online review are less complicated and provides more detailed result as compared to tweets. Build a dictionary for tweets are complicated as it includes internet slang, sarcasm. So, social media are hard to analyse as they have their unique structure and grammar.

2) Mathew Smith, Christian Szongott, Benjamin Henne, Gabriele von Voigt. "***Big data privacy issues in public social media***",**(2013)** stated how the capabilities of mobile devices are affecting user's privacy. It also presents threat analysis which is classified into two categories i.e. home grown problem in which user upload without sufficient protection which affect user's own privacy. Second, someone is uploading the damage content of other people. It also include privacy analysis of different sites such as flicker, Face book, Picasa web and Google+, Locr and Instagram and PicPlz. It also presents an analysis of privacy related metadata, particularly location data contain in social media. As it concludes that 10% of all the photos taken by camera devices harm other people's privacy without knowing them. It also represents handling of the location based big data. It includes a concept of watchdog client a server side watchdog service. In it, concept to stay in control from social media uploaded by others has three types of services. Through the regular user account, Operated by the social networks and last one operated by third party i.e. stand alone service.

3) Javier Conejero,Peter Burnap,Omer Rana,Jeffrey Morgan. "***Scaling Archived Social media Data Analysis using a hadoop cloud***", **(2013)** paper presents a COSMOS platform for sentiment and tension analysis on twitter dataset. Tool used for sentiment analysis is SeniStrength. To run application based on cloud environment, it uses virtualized Hadoop Clusters in OpenNebula. This system configuration used for performance aspects which shows how virtual server needs to be distributed as to reduce variability in the analysis performance. It also presents the architecture for data processing of COSMOS using OpenNebula and Hadoop. Processing performance comparison is done over Cardiff Cloud Tweet and UCLM Cloud Tweet which shows Cardiff Cloud have better performance due to its compute node has been more powerful than UCLM Cloud compute node. This paper involves future work to evaluate on bigger cloud environment and increase number of virtual cluster and Twitter message and improve performance with multiple concurrent users using the same cloud service. As using COSMOS we can add more nodes and workers to the problem and bring processing time down further.

4) Ya-Ting Chang, Shih-Wei Sun. "***A Real-time Interactive Visualization System for Knowledge Transfer from Social Media in a Big Data***", **(2013)** stated a proposed real-time interactive visualization system. This paper is contributed towards three objectives a) analyze and visualize system from a social media on a real time basis. b) Kinect camera and a mobile device is used to interact with the system. c) Knowledge has been transferred from social media in big data providing Geo-location of social media, suggesting the path

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 2, Issue 1, ISSN No. 2455-2143, Pages 67-72
Published Online November-December 2016 in IJEAST (http://www.ijeast.com)

or route and then generating images. A proposed real time system consist of three parts i.e.

Analysis and visualization –here data are collected from social media and sent for analysis purpose. For analyzing partnership of a social media it has used Node XL.

i) A kinect camera is used to track a user. Movements of user are tracked and are displayed in Virtual Reality environment. For the identity purpose, once user login in proposed system their identity has been recorded by the user: fusing sensors on a mobile device and hand joint.

ii) Shows the relationship between users and related multimedia content at that time. This paper provides a comparison on social network relationship, location based service and data visualizing.

5) Simona Vinerean, IulianaCetina. "*The Effects of Social Media Marketing on Online Consumer Behavior*", (2013) stated to answer the question of how people interact on online and how they are engaged in online activities. Study include online activities of 236 Social media users, by identifying different types of users, a segmentation of these users and a linear model is designed to examine how different predictors are related to social networking sites that consider a positive impact on the respondents perception of online advertisements. This study can help to discover how to engage different types of audience in order to maximize the effect of the online marketing strategy. Limitation of study include with online questionnaires, which include unsystematic sampling procedures and low response rate. Future research can be measured based on demographic variables such as sex, age and social class.

III. PRESENT WORK

3.1 Problem Formulation

Different types of data are generated from different Social media groups that need to be organized and to monitor people's attitude towards products, gadgets, movie review etc. This database is collected from different social media sites for example Twitter, Facebook, Online review, shopping sites etc. Text analytics and Sentiment analysis can help to develop valuable business insights from text based contents that may be in the form of word documents, tweets, comments and news that related to Social media. The foremost reason of Sentiment analysis is so complex is that words often take different meanings and are associated with different emotions depending on the domain in which they are being used. Dataset is analyzed by using the weka tool. The hidden relationship has to be extracted from this type of database using different mining approaches in Weka tool. Dictionary building for detailed sentiment analysis implies making an initial list of adjectives and nouns which are normally used when describing a specific movie review. Phrases and terms are extracted from this relational dataset and their meaning has been added to dictionary for next generation analysis. In tweets, informal and shortcuts has been used for explaining terms or views and this is done with the help of sentiments analysis is not an easy process. To reduce this, data mining approaches has been used for extraction of features from these datasets.

3.2 Objective

Our objective is to collect and preprocessing of raw data for movies reviews and to filter the data using string to word vector. To improve Naïve Bayes Classification algorithm on collected data. Than, we analyze the performance and compare it with the existing algorithm. We also construct combined dictionary from online review and twitter dataset keywords i.e; from movie reviews.

3.3 Methodology

1) Collection of raw data and then apply filtering techniques to make that raw data into structured format. For doing the classification, Text preprocessing and feature extraction is a preliminary phase. Preprocessing involves 3 steps:

a) **Word parsing and tokenization:** In this phase, each user review splits into words of any natural processing language. As movie review contains block of character which are referred to as token.

b) **Removal of stop words:** Stop words are the words that contain little information so needed to be removed. As by removing them, performance increases. Here, we made a list of around 320 words and created a text file for it. So, at the time of preprocessing we have concluded this stop word so all the words are removed from our dataset i.e. filtered.

c) **Stemming:** It is defined as a process to reduce the derived words to their original word stem. For example, "talked", "talking", "talks" as based on the root word "talk". We have used Snowball stemmer to reduce the derived word to

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 2, Issue 1, ISSN No. 2455-2143, Pages 67-72
Published Online November-December 2016 in IJEAST (http://www.ijeast.com)

their origin.

2) Apply the improved Naïve Bayes algorithm for classification.

1. Combining naïve bayes with Decision table using Decision tree as Meta classifier.

2. Meta Learner is a learner scheme that combines the output of the naive bayes and decision table i.e. the base learner. The base learners' level-0 models and the meta-learner is a level-1 model. The predictions of the base learners are input to the meta-learner.

3) Analyze the performance parameters like FP rate, TP rate, Recall, Precision of Naïve Bayes and new proposed hybridized algorithm and Compare the results of both.

Now for evaluating the result, different parameter are to be calculated. True positive, True negative, False positive and False negative are used for comparing the class label that have been assigned to a document by the classifier with the classes the item actually belongs [18].

a) **Accuracy**: It is measured as the proportion of correctly classified instances to the total number of instances being evaluated. Classification performance being evaluated by using this parameter. where True positive – that are truly classified as positive.
False positive- not labeled by the classifier as positive but should be True negative- that are truly classified as negative
False negative- not labeled by the classifier as negative but should be
b) **Precision**: It is widely used in evaluating the performance in different field such as text mining, information retrieval. Precision is also referred to measure the exactness. It is defined as ratio of the number of correctly labeled as positive to the total numberthat has been classified as positive.
c) **Recall**: It is also used in evaluating the performance for text mining and information retrieval. It is also used to measure the completeness of the model. It is defined as the ratio of the number of correctly labeled as positive to the total number that are truly positive.
d) **F-measure**: It is referred as the harmonic mean of precision and recall. It helps to give score needed to balance between precision and recall. It combines two of them into one for the convenience as it might optimize the system so that it can favor one of them.

4) Combined dictionary

Combined word of twitter dataset and online review dataset forms a dictionary. As after classifying each word probability as positive, negative and neutral. Compare the probability for each word and categorize each word into three different dictionaries based on highest polarity i.e. positive, negative and neutral of each word. Dataset is used for further evolution of words depending on their uses in daily life as adjectives or nouns in the social media data.

FLOWCHART OF PROPOSED METHODOLOGY

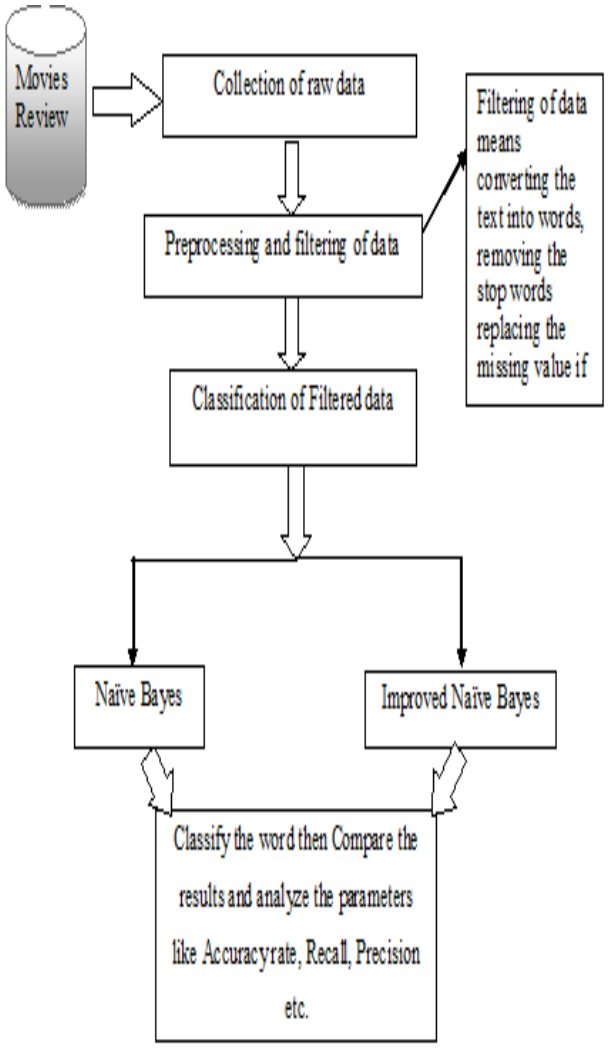

Fig 1- Flowchart Of Proposed Methodology

IV. CONCLUSION

Social media Monitoring has been growing very rapidly so there is a need for various organizations to analyze customer behavior or attitude of particular product or any movie review. So, the concepts of sentiment analysis have been introduced. Text analytics and sentiment analysis can help organization to derive valuable business insights. Attitude can be calculated based on polarity check. Sentiment analysis refers to a broad range of fields of natural language processing, computational linguistics, and text mining. Sentiment classification of reviews and comments has merged as the most useful application in the area of sentiment analysis. Bag of words and feature based sentiment are the most popular approaches used by researchers to deal with sentiment analysis of opinions about products such as movies etc. Sentiment analysis on movies review are done by forming dictionary which shows that it is easier to build dictionary on phrases of movies reviews. In this, level sentiment analysis is considering three classes for sentiment polarity of each sentence (positive, neutral and negative). Each class prediction and classification is done by algorithm in terms of accuracy, precision, recall etc. Also the comparison of Naïve bayes with Improved Naives Bayes is done on the basis of accuracy or the correctly classified instances.

V. REFERENCES

[1] Ana Mihanovic, HrvojeGabelica, ZivkoKrstic (2014) "*Big Data and Sentiment Analysis using Knime: Online Reviews Vs. Social Media*", MIPRO Opatija, Croatia.

[2] BogdonBatrinca, Philip C. Treleaven (2014) "*Social media analytics: a survey of techniques, tools and platform*" Department of Computer Science, Gower Street, London, UK published in Springer.

[3] Matthew Smith, Christian Szongott, Benjamin Henne, Gabriele von Voigt (2013) "*Big Data Privacy Issues in Public Social Media*", Distributed Computing & Security Group, Leibniz Universitat Hannover, Thailand, Germany IEEE.

[4] Javier Conejero, Peter Burnap, Omer Rana, Jeffery Morgan (2013) "*Scaling Archied Social Media Data Analysis Using a Hadoop Cloud*" sixth international conference on cloud computing, IEEE.

[5] Ya-Ting Chang, Shih-Wei Sun (2013) "*A Real time Interactive Visualization System for Knowledge Transfer from Social Media in a Big Data*", Center for Art and Technology, Taipei National University of the Arts, Taipei, Taiwan, IEEE.

[6] NargizaBekmamedova, Graeme Shanks (2013) "*Social Media Analytics and Business Value: A Theoretical Framework and Case Study*", Department of Computing and Information Systems, University of Melbourne.

[7] Simona Vinerean, IulianaCetina (2013)"*The Effects of Social Media Marketing on Online Consumer Behavior*",International Journal of Business and Management; Vol. 8, No. 14.

[8] Kristin Glass and Richard Colbaugh (2012) "*Estimating the sentiment of social media content for security informatics applications*", Institute for Complex Additive System Analysis, Socorro, USA, Springer Journal.

[9] SitaramAsur, Bernardo A.Huberma (2012) "*Predicting the Future with Social Media*",Social Computing Lab, HP Labs, Palo Alto, California.

[10] A.Iosup, N.Yigitbasi and D.H.J Epema (2011),"on the performance variability of production cloud services", in proceedings of CCGRID.

[11] LingyanJi, Hanxiao Shi, Mengli, MengxiaCai, PeiqiFeng. (2010) "Opinion Mining of Product reviews based on semantic role labeling", 5th International Conference on Computer Science and Education, IEEE, pp. 1450-1453, August 2010.

[12] Mohsen Farhadloo, Erik Rolland. (2013) "Multi-class Sentiment analysis with clustering and score representation", 13th International Conference on Data mining Worshops, IEEE, pp. 904-912, December 2013.

[13] Mrs. R.Nithya, Dr.D.Maheshwari. (2014) "Sentiment Analysis on Unstructured Review", International Conference on Intelligent Computing Application, IEEE, pp. 367-371, March 2014.

[14] Ms. K. Mouthami, Ms. K.Nirmala Devi, Dr.V.MuraliBhaskaran. (2010) "Sentiment Analysis and Classification based on Textual Reviews",Dept of CSE, Tamil Nadu, IEEE.

[15] NargizaBekmamedova, Graeme Shanks (2013) "Social Media Analytics and Business Value: A Theoretical Framework and Case Study", 2014 47th Hawaii International Conference on System Sciences (HICSS), pp. 3728-3737, January 2014.

[16] SimonaVinerean, IulianaCetina (2013) "The Effects of Social Media Marketing on Online

Consumer Behavior", International Journal of Business and Management; Vol. 8, No. 14.

[17] SitaramAsur, Bernardo A.Huberma (2010) "Predicting the Future with Social Media",IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), Vol. 1, pp. 492-499, 2010.

[18] V. S. Jagtap, KarishmaPawar. (2013) "Analysis of different approaches to Sentence-Level Sentiment Classification", International Journal of Scientific Engineering and Technology, Vol. 2, Issue 3, pp. 164-170, April 2013.