ISSN NO. 2320-5407

Journal homepage: http://www.journalijar.com
Journal DOI: 10.21474/IJAR01

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH

RESEARCH ARTICLE

SENTIMENT ANALYSIS OF SOCIAL MEDIA DATA USING NAIVE BAYESIAN CLASSIFIER IN HADOOP AND HIVE.

G.Mani, G.Jyothi, G.Swathi, Ravuri Daniel.
Department of Information Technology, Vignan's Institute of Information Technology, Visakhapatnam, India.

Manuscript Info

Manuscript History:

Received: 18 February 2016
Final Accepted: 22 April 2016
Published Online: May 2016

Key words:
Sentiment,Polarity,Twitter,HDFS,Hive.

Corresponding Author
G.Jyothi.

Abstract

Sentiment Analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. Tweets are frequently used to express a tweeter's emotion on a particular subject. There are firms which poll twitter for analyzing sentiment on a particular topic. The challenge is to gather all such relevant data, detect and summarize the overall sentiment on a topic. The social media data includes the collection of unstructured raw data from various social media posts & blogs such a Twitter and needs to refine the data to get the results which are most important for understanding the sentiment of a product or service is either positive, negative or neutral. The steps for extracting sentiment data from Twitter and analyzing the performance of a recent iron man3 movie release. we can mine twitter, Facebook and other social media conversations for sentiment data about a company products or movies etc is used to make targeted, real-time, decisions that increase market share. In our project we proposed an architecture, where the raw data taken from twitter is classified and added to HDFS using Naïve Bayesian classifier. Further we process HIVE queries on the sentimental data to catalogue positive, negative and neutral tweets. Subsequently the analysis is illustrated using BI tools. The main advantage of our project is we can visualize divergent tweets on the given attributes and data set according to one's choice in a map view.

Introduction:-

In recent years it has been observed that micro blogging website twitter have progressed to become a source of various kind of information and playing an important role for sentiment analysis world-wide. According to statistics of twitter user (Twitter usage stats) more than 600 million tweets are tweeted every day and this numbers are increasing. Analyzing tweets has applications in understanding how public sentiment is shaped, tracking public sentiment and polarization with respect to their views and issues and understanding the impact of tweets from various entities. Sentiment analysis helps them to take decision about their future projects.

In this paper, We Extract, Transform and utilize unstructured data generated by social networks respectively. The generated data is then analyzed through Naïve Bayes classifier display the positive, negative tweets classification in dictionary file and processing of data by using Hadoop and Hive .This analysis will be shown with interactive visualizations using some powerful BI tools for Excel like Power View. Finally, a real time case study will be used to create a report on how Sentiment Analysis can be implemented for classification of ironman3 movie review tweets according to polarity. The main advantage is we can visualize the tweets country wide based on the given attributes and data set which contains unstructured data taken from twitter. The application inputs the dataset that contains the information. By using HIVE, a component in HADOOP we write queries on the input datasets according to the users perception and generate the results accordingly in a fast and accurate mode.

The remaining part of the paper organized as follows. The literature review of the proposed work described in Section II,Section III explains the proposed system architecture and methodology of Naïve Bayes classification. The results and discussion are explained in Section IV and we conclude conclusion in Section V.

Literature review

Varsha Sahayak, Vijaya Shete, Apashabi Pathan [10], Now-a-days social networking sites are increasing huge amounts of data i.e. generated every day. Millions of people are sharing their views every day on micro blogging sites, since it contains short and simple expressions. Here the paradigm to extract the sentiment from a famous micro blogging service, Twitter, where users post their opinions for everything. In this paper previously twitter dataset is analyzed by data mining approach such as use of Sentiment analysis algorithm using machine learning algorithms. An approach is introduced that automatically classifies the sentiments of Tweets taken from Twitter dataset as in. These messages or tweets are classified as positive, negative or neutral with respect to a query term. This is very useful for the companies who want to know the feedback about their product brands or the customers who want to search the opinion from others about product before purchase. They use machine learning algorithms for classifying the sentiment of Twitter messages using distant supervision which is discussed in .the training data consists of Twitter messages with emoticons, acronyms which are used as deafening labels discussed in [2]. We examine sentiment analysis on Twitter data. The contributions of this survey paper are: (1) we use Parts Of Speech (POS)-specific prior polarity features. (2) We also use a tree kernel to prevent the need for monotonous feature engineering.

Sunil B. Mane, Yashwant Sawant, Saif Kazi, Vaibhav Shinde[7],Twitter, one of the biggest social media site receives tweets in millions every day. This large amount of raw data can be used for industrial or business purpose by establishing according to our requirement and processing. This paper provides a way of sentiment analysis using Hadoop which will develop the large amount of data on a Hadoop cluster faster in real time. Here they focused more on the speed of performing analysis than its accuracy i.e. performing sentiment analysis on big data which is completed by splitting the various modules of data in following steps and working together with Hadoop for planning it onto different machines. Part of speech tagged using openly. This tagging is used for following various purposes.

- Stop words removal:. All the words having this tag are not considered. The stop words like a, an this which are not useful in performing the sentiment analysis are removed in this phase.
- Unstructured to structured: Twitter comments are mostly unstructured i.e. ‘Sry’ is written ‘sorry’, ‘tc’ to actually ‘take care’. Conversion to structured is done by dynamic data records of unstructured to structured and vowels adding.
- Emoticons: These are most expressive method available for opinion. The emoticons figurative representation is converted in to words at this stage i.e. to joyful.

Penchalaiah.C,Murali.G,Suresh Babu [4], Here we have categorized this sentiment analysis into 3 groups like tweets that are having positive, moderate and negative comments. As of now we know present businesses and some survey firms are mainly taking decisions by data obtained from blogs. Here we are going to talk how effectively sentiment analysis is done on the data which is collected from the Twitter using Flume. We can collect the data from the twitter by using BIGDATA ecosystem using online streaming tool Flume. Twitter is an online web application which contains large amount of data that can be a structured, semi structured and unstructured data. So here we are taking sentiment analysis, for this we are using Hive and its queries to give the sentiment data based up on the groups that we have defined in the HQL (Hive Query Language).

Apoorv Agarwal Jasneet Singh Sabharwal [1][2], we compare the performance of this hierarchal pipeline with that of a 4-way classification scheme. However, to the best of our knowledge, there is no work that explores the classifier design matters explored in this paper. In addition, we explore the trade-off between making a prediction on lesser number of tweets versus F1-measure. We present an end-to-end pipeline for sentiment analysis of a popular social media site called Twitter. We build a hierarchal cascaded pipeline of three models to label a tweet as one of Objective, Neutral, Positive, Negative class. Generally we show that a cascaded design is better than a 4-way classifier design.

Proposed System Architecture and Methodology:-

System architecture

In the Hadoop ecosystem, the proposed system architecture is shown in Fig 1 the Hive application provides a query interface which can be used to query data that resides in HDFS. The data is taken from database and processed to naïve Bayesian classifier and further processed to MapReduce. Then we can query the twitter raw data to get the result of the refined sentiment data from which we can able to generate reports and analyze the Twitter Sentiment Data. These result data is downloaded and we access the refined sentiment data with excel to visualize the sentiment data using excel power view according to country wide sentiment analysis us to easily model complex types, so we can easily query the type of data. Further it is categorized into positive tweets and negative tweets. Then visualization is done.

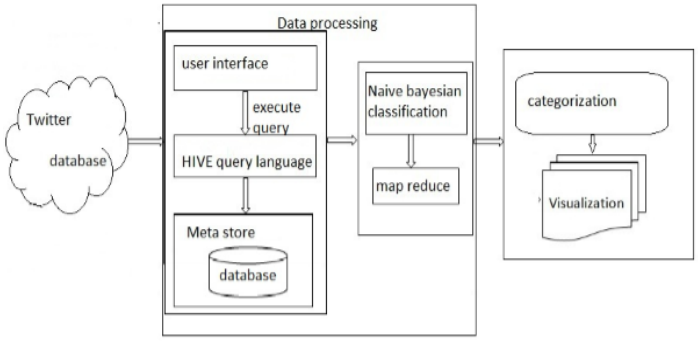


Fig 1: Proposed system Architecture.

In the proposed system architecture the application inputs dataset that contains the information taken by the twitter which is in unstructured format. The raw data taken from twitter is classified and added to HDFS using Naïve Bayesian classifier. Further we process data by mapreduce jobs by giving Hive queries on the sentimental data to catalogue positive, negative and neutral tweets. In Hadoop and Hive by writing Hive queries we process the data into tables. The result data in structured format shown in tables will be downloaded in csv file. By using BI tools , we visualize the results in power view, which display the sentiment analysis of data according to country in map view.

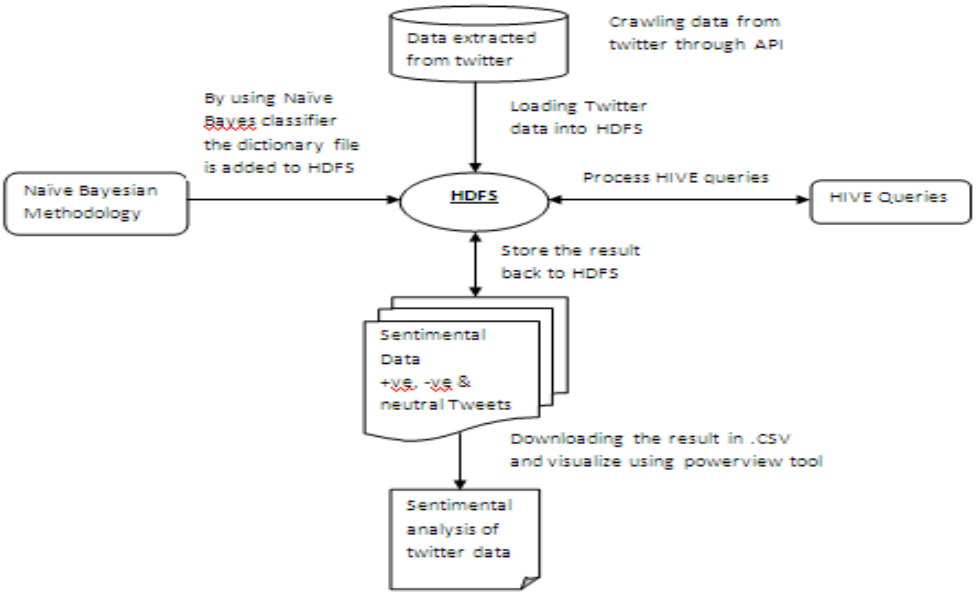


Fig 2: Work flow process of system architecture.

- Crawling twitter data set.
- Loading these dataset, and classify the positive and negative keywords using Naïve Bayesian Classifier and stores that dictionary file into HDFS.
- Then, implementing the HIVE queries through HIVE script and stores the result back to HDFS.
- Then according to the query sentiment analysis is done here.
- Then, results of sentiment data will be downloaded in CSV file to visualize the results in map view according to country wide.

Methodology:-

In our proposed system we use the Naive Bayesian classifier i.e. based on Bayes theorem with independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods. Naive Bayes is Bayesian probability distribution model based algorithm. In general all Bayesian models are derivatives of the well-known Bayes Rule, which suggests that the probability of a hypothesis given a certain evidence, i.e. the posterior probability of a hypothesis, can be obtained in terms of the prior probability of the evidence, the prior probability of the hypothesis and the conditional probability of the evidence given the hypothesis. Mathematically,

$$P(H/E) = P(H)P(E/H)/(P(E)) \quad (1)$$

Where,

P(H|E) - posterior probability of the hypothesis.

P(H) - prior probability of hypothesis.

P(E) - prior probability of evidence.

P(E|H)-conditional probability of evidence of given hypothesis.

Or in a simpler form:-

$$\text{Posterior} = ((\text{Prior}) \times (\text{Likelihood}))/\text{Evidence} \quad (2)$$

To explain the concept, lets take an example. For instance, we have a new tweet to be classified in to one of the positive or negative classes. Given that in the previously classified tweets, positive tweets are twice the number of negative tweets. Since the new tweet's class is not known, the problem is estimating correctly the class that the tweet is to be categorized in. This can be found out by Bayes rule calculating the probabilities of the likelihood of the tweet to be positive or negative. Hence, Design and Analysis of Proposed Approach we have:

$$P(n/p) = \frac{p(n)P\left(\frac{p}{n}\right)}{P(p)} \quad (3)$$

Since there are twice as many positive tweets as negative, it is reasonable to believe that a new case (which hasn't been observed yet) is twice as likely to have membership positive rather than negative. In the Bayesian analysis this belief is known as the prior probability. Prior probabilities are based on previous experience. In this case the percentage of positive tweets and negative tweets, and often used to predict outcomes before they actually happen. Thus, we can write:

Prior Probability of positive tweet:

$$P(p) = \frac{\text{No.of positive tweets}}{\text{Total no.of tweets}} \quad (4)$$

Prior Probability of negative tweet:

$$P(n) = \frac{\text{No.of negative tweets}}{\text{Total no.of tweets}} \quad (5)$$

Let there be say a total of 6k tweets, 4k of which are positive and 2k negative, our prior probabilities for class membership.

Prior Probability for positive tweet P(p) = 6k/ 9k = 6/9 = 2/3

Prior Probability for negative tweet P(n)= 2k /6k = 2/ 6 = 1/ 3

The likelihood of the tweet falling into either of the classes is equal, since we have only two classes. So likelihood of X = 0.5. So now calculating the posterior probability of the new tweet say X, being positive or negative will be:

•Posterior probability of X being positive = (Prior probability of positive)× (Likelihood of X being positive) = 6/9 × 1/2 = 1/3 = 33.34% chances of X being positive.

•Posterior probability of X being negative = (Prior probability of negative)× (Likelihood of X being negative) = 1/3 × 1/2 = 1/6 = 16.67% chances of X being negative.

Thus this tweet will fall in to the positive class:-
In our case we would have two hypothesis and many other features on basis of which the one that has the highest probability would be chosen as a class of the tweet those sentiment is being predicted. After every classification step all the probabilities are again calculated and updated accordingly. Sentiment Analysis is automatic extraction of subjective content of text and predicting the subjectivity such as positive or negative.

Sentiment Analysis is the process of taking a block of text and determining if the author feels positive, neutral, or negative about a particular topic. It can be an extremely difficult problem to do correctly. For instance, consider the following (naive) approach. This approach simply takes a dictionary of words and assigns a positive or negative weight to each. To determine the overall sentiment of a phrase, simply add up the scores of the words found in the dictionary set are shown below:

Table 1: The example for keyword and sentiment.

Keyword	**Sentiment**
Excellent	10
Impressed	6
Great	5
Ok	1
Meh	0
Boring	-3
Sick	-4
Terrible	-5

Now, taking this table, we can assign values to the following sentences:
- The movie was great! Excellent explosions! 5 + 10 = 15
- I thought the movie was terrible. Boring! -5 + (-3) = -8

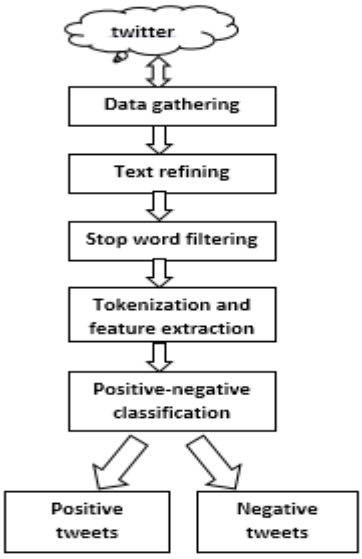


Fig 3:- Process for classification of tweets.

The above Fig 3 shows that we gather information from twitter and process data by using Naïve Bayesian classifier to divide the bunch of words into positive and negative tweets. These words form into a type of dictionary set, and that data is processed by map reduce algorithm to give more accuracy and better performance as shown below tables.

Table 2: Performance of naive Bayes classifier

Dataset size	Accuracy
1K	63.95
10K	97.50
100K	95.48
500K	98.60

The performance of Naïve Bayesian classifier with dataset size, and finding its accuracy is the ability of the classifier to correctly classify and label the new tweets into their respective classes. The comparison of classifier performances with dataset size (no of tweets), and finding its accuracy. Accuracy is the ability of the classifier to correctly classify and label the new tweets into their respective classes. In the above figure we compare the old and new accuracy results according to dataset size.

Results and Discussion:-

The performance of the proposed algorithm in Hadoop is accurate and processing time is less compared with different Traditional data mining tools and techniques. The performance evaluation of data in Hadoop is shown in below table. In this the CPU cumulative time is very less in processing the data while compared with traditional data mining tools. Then data processed by the Hive queries in hive editor. The hive editor contains the results information, query, logs that contains the cpu cumulative time and also the columns of the dataset which contains in the data. The process related information while the data is processed like CPU cumulative time, no of nodes created and the no of mapper and reducer jobs are created. Let's use HCatalog to take a quick look at the data. Open the Sandbox HUE user interface, then click HCatalog in the menu at the top of the page.

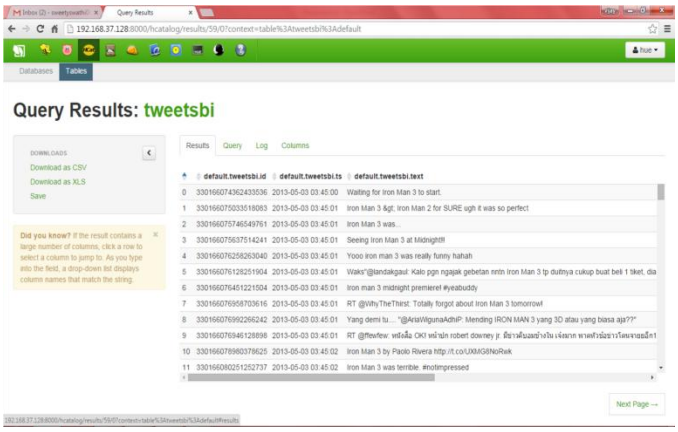
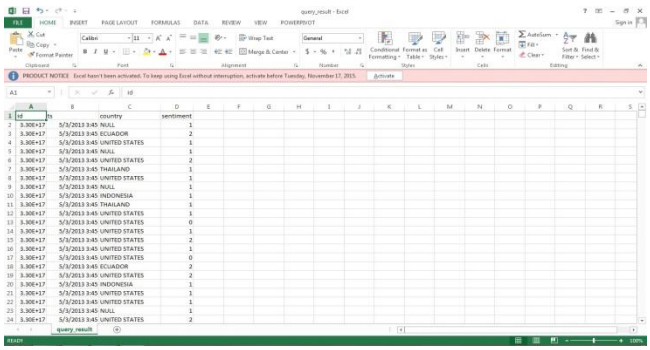


Fig 5:- Resulting the Query Of Tweets bi.

In the above Fig 5 which resulting the query of tweets bi displays the sentiment analysis of tweets according to the polarity. By selecting the tweets bi folder and click on upload.

The "tweetsbi" table is the table created by the Hive script that added a column with the sentiment value for each tweet. Now that we have refined Twitter data in a tabular format with sentiment ratings, we can access the data with Excel. After removing the column tweets. The imported query data appears in the excel workbook.



Now that we have successfully imported the Twitter sentiment data into Microsoft Excel, we can use the In the Excel worksheet with the imported "tweetsbi" table, select Insert Power View to open a new Power View report. The Power View Fields are appears on the right side of the window, with the data table displayed on the left. In the Power View Fields area. Then clear the checkboxes next to the id and ts fields, then click Map on the Design tab in the top menu. Now let's display the sentiment data by color. In the Power View Fields area, click sentiment, then select Add as Color. The map displays the sentiment data by color:

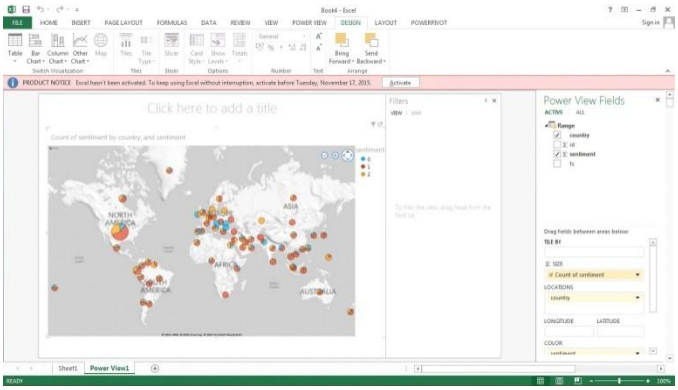


Fig 6:- The map displays the sentimental data by color.

By analyzing the dataset, which is shown in Fig 6. we obtained result of sentiment analysis of positive, negative and neutral tweets of IronMan3 Movie review according to country wide:

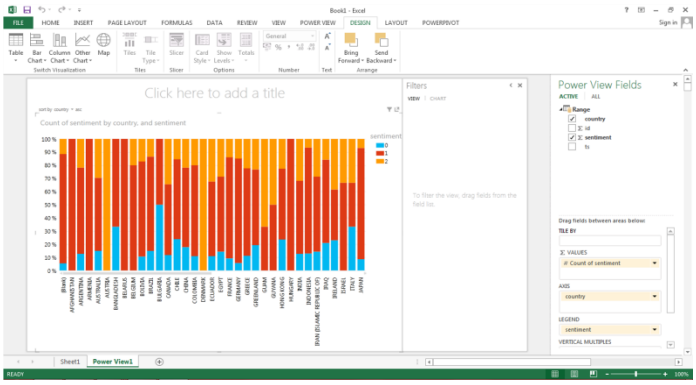


Fig 7:- Displays the sentiment data in bar graph.

The above Fig [7] displays the sentiment data in bar graph according to country wide as blue color indicates the negative tweets. The orange color indicates the positive tweets, and red color indicates the tweets are neutral.

Table 3: Sentimental Analysis of Country Wide Tweets.

S.no	Positive Tweets	Negative tweets	neutral tweets	Country
1	14%	15%	60%	Brazil
2	30%	19%	65%	China
3	22%	23%	52%	hongkong
4	5%	10%	85%	Japan
5	100%	nill	nill	Australia

Here the sentiment analysis of country wide tweets table[3] are shown, by above Fig [7]. Where the above fig displays the sentimental analysis in bar graph according to worldwide statistics. Then some of countries are taken as example which is displayed in above table.

Table 4: Performance Evaluation table.

Size of data	Total input paths to process	No of splits	No of mappers	No of Reducers	Cumulative CPU time
5234 kilobytes	29	29	29	1	191.21sec

The performance of the proposed system is tested in our dataset shown in Table [4]. Which contains of size 5234 kilo bytes is executed and process using the Hive queries. After performing these queries the process is done by MapReduce jobs. Then after processing these logs the cumulative CPU time is 191.21 sec.

Conclusion:-

Today's world is more and more rely on consumer experience and opinions of a service. In this project we have shown a visualization of new movie review using statistics of the tweets labels that predicts the movie rating (hit/flop/average).The sentiment analysis of tweets about a movie makes one aware about different aspects of the movie and plays a major role in decision making. This information will be useful for planning marketing activities for any future movie releases or any other company products.

Future work:-

- Sentiment Analysis can be very effective in predicting Election results, stock market or movie review.
- Like Imdb (internet movie database) reviews of Facebook and twitter can be also used to give useful data which can be used to predict future.

References:-

1. Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau, "Sentiment Analysis of Twitter Data" Department of Computer Science, Columbia University, New York, 2009.
2. Apoorv Agarwal Jasneet Singh Sabharwal "End-to-End Sentiment Analysis of Twitter Data".
3. Efthymios Kouloumpis, TheresaWilson, Johanna Moore Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media
4. Penchalaiah.C,Murali.G,Suresh Babu.A "Effective Sentiment Analysis on Twitter Data using Apache Flume and Hive" IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 8, October 2014.
5. Puneet Singh Duggal ,Sanchita Paul "Big Data Analysis: Challenges and Solutions" International Conference on Cloud, Big Data and Trust 2013, Nov 13-15 , RGPV
6. SanjeevDhawan, Sanjay Rathee " Big Data Analytics using Hadoop Components like Pig and Hive"gef
7. Sunil B. Mane, Yashwant Sawant, Saif Kazi, Vaibhav Shinde "Real Time Sentiment Analysis of Twitter Data Using Hadoop" Sunil B. Mane et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 3098 – 3100
8. Theresa Wilson, Joanna Moore, Efthymios Kouloumpis, "Twitter Sentiment Analysis – The Good, the Bad and the OMG" Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media .
9. T.K.Das1 ,P.MohanKumar,"BIG Data Analytics: A Framework for Unstructured Data Analysis"T.K.Das et al. International Journal of Engineering and Technology (IJET).
10. Varsha Sahayak ,Vijaya Shete, Apashabi Pathan "Sentiment Analysis on Twitter Data", International Journal of Innovative Research in Advanced Engineering (IJIRAE) 163 Issue 1, Volume 2 (January 2015)