

Sentiment and Predictive Analysis of Big Data for Hotel Reviews

Author: Priti Gupta¹; Arvind Upadhyay²

Affiliation: PG Scholar, Computer Science & Engineering Department, IPS Academy, Indore (M.P.), India¹; Assistant Professor, Computer Science & Engineering Department, IPS Academy, Indore (M.P.), India²

E-mail: pritigupta.061189@gmail.com; upadhyayarvind10@gmail.com

ABSTRACT

At present, with the volume of data growing at an exceptional rate, big data mining and information finding have developed into a narrative challenge. Scientists and computer engineers have coined a new term for the phenomenon: "Big Data". "Big Data" involves using multiple data sources, internally and externally. Automated analysis is required in Science projects with the growing availability and popularity of feedback reviews resources such as online review sites and personal blogs. New opportunities and challenges arise as people now can, and do, actively use information technologies to seek out and understand the opinions of others. The sudden eruption of activity in the area of opinion mining and sentiment analysis, which deals with the computational treatment of opinion, sentiment, and subjectivity in text, has thus occurred at least in part as a direct response to the surge of interest in new systems that deal directly with opinions as a first-class object. This analysis covers techniques and approaches that promise to directly enable opinion-oriented information-seeking systems for user provided reviews data which are available in vast amount. In this research, we propose the fusion prediction model for data access pattern. The goal is to predict the next pieces of data needed in the processor and preload them into the memory in order to improve the overall processing time. Our proposed model can deliver high prediction accuracy. The now introduced MapReduce technique has received much attention from both scientific community and industry for its applicability in big data analysis.

Key Words: Big Data, Sentiment Analysis, Predictive Analysis, Opinion Information, Users-Reviews.

I. INTRODUCTION

Big Data has become one of the buzzwords in IT during the last couple of years. Big data is completely the rage in the commercial world these days, and insurance which exists and respire by quantities has hesitantly hopped on the movement. But what precisely is big almost the new sources of data in a coverage context and how it dissimilar from completely of the other information is that prerogatives manager have traditionally tapped into throughout their investigations. Let us revenue an instant to distinct the hype from the genuineness. Is there actually a mountain of gold information in big data mines what way strength claims managers most efficiently dig up and improve these novel, possibly valuable data resources First of every one, there is a lot further data out there for insurers to slice and dice together organized as well as unstructured (meaning raw data that necessity be sifted finished and sophisticated to have several real value). Supervision data volume and rapidity can be difficult, like demanding to quench your thirst by drinking out of a fire hose. Indeed, some carriers are having trouble correlating all the internal data they already generate, so it might be wise for insurers to create certain they have their own processing houses in order before flooding the pipeline with additional material.

But big data management should be achievable for insurers. After all, carriers have a lot more volume at their discarding. Memory is inexpensive, Processing is faster. Investigative programs are progressively refined. But recall that the benefit of big data is not basically having added information. Certainly, congregation data should not be an end in themselves insurers also essential to be clever to attach these new data streams to division actionable outcomes.

Big Data is not just about the size of data but also includes data variety and data velocity. Together, these five attributes of Big Data.

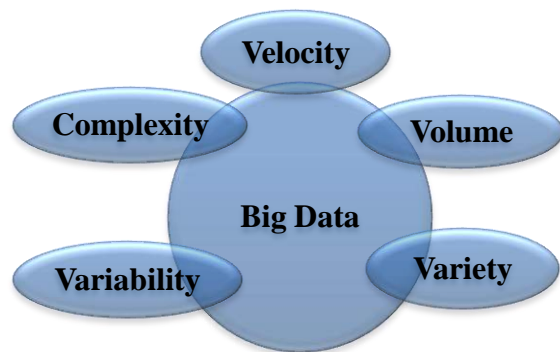


Figure 1: The Characteristics of Big Data

The main research questions are related but not limited to the following aspects:

- In the context of big data and cloud computing how analytics information and knowledge management disciplines and approaches will develop.
- What should be the technique, strategy and practices to increase the benefits and minimize the big data risks?
- A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level — whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral. Advanced, "beyond polarity" sentiment classification looks, for instance, at emotional states such as "angry," "sad," and "happy".

- It difficult to identify nouns and adjectives that can be classified as sentiments. The next step is to filter what matters. Remove non-textual contents and mark-up tags and other data that is not required for the analysis.
- The obtainable legislation such as data protection law, system and standards how should develop. Furthermore, the ethics issues will be considered that could be easier supposed than complete with big data, since exporters may have to pan finished a lot of raw materials to discover the few gold nuggets that could meaningfully impact a specific rights study.

Nowadays, people are extensively utilizing various social media communication means: they create online accounts in social networks to keep in touch with friends and business partners, run and read blogs in order to share opinions, obtain important information, advertise and sell their products online.

User provide review based on their experience, so using that reviews data we will provide sentiment on that reviews means whether user has given positive review about thing or negative and also predict score/rank by using that review. It's more of Sentiment analysis and less predictive analysis as we are predicting score/rank based on the reviews.

Let's take example of Hotels/Restaurants, whenever user's uses some hotel and restaurant they provide there review based on their experience in hotel, like room quality, hotel services, food quality, value for money hotel etc. So based on that review we have to determine whether that review is positive or negative and we can also determine ranking of hotels.

The simplicity and low-cost of the usage of communication means allows customers to publish their experiences and opinions about the purchased products and services online by creating a blog post or giving a positive or negative assessment in products' review portals.

II. RELATED WORK

Several good surveys of various ensemble learning techniques can be found in [11], including voting [17], bagging [4], boosting [14], stacked generalization and cascading [8] [5], etc. All these methods share the same principle: each local learner trains its own learning model independently, which requires no communication with other local learners at all. Hence, local learners are trained in a non-cooperative way which leaves unexplored the hidden rules that inform how different local learners are correlated and how they should be coordinated in training. In contrast to such non-cooperative methods, cooperative distributed data mining techniques seek to improve prediction accuracy at the price of some mild communication costs between local learners.

V.D.Nguyen et al. [1] proposed the two underlying approaches for sentiment analyses are dictionary based and machine learning. The former is popular for public sentiment analysis, and the latter has found limited use for aggregating public sentiment from Twitter data. This paper aims to extend the machine learning approach for aggregating public sentiment. To this end, a framework for analyzing and visualizing public sentiment from a Twitter corpus is developed.

N. D. Valakunde et al. [12] proposed approach allows us to automatically take care of the current problems of document level sentiment analysis, such as, entity identification, subjectivity detection and negation. The technique is further applied for educational data mining, where a faculty performance is evaluated using the sentiment analysis of comments provided by students as a part of their feedback. The document level faculty performance score is computed from the distinct aspect based sentiment scores, such as, knowledge, presentation, communication and regularity of the faculty.

G.Vinodhini et al. [9] observed accurate method for predicting sentiments could enable us, to extract opinions from the internet and predict online customer's preferences, which could prove valuable for economic or marketing research. Till now, there are few different problems predominating in this research community, namely, sentiment classification, feature based classification and

handling negations. This paper presents a survey covering the techniques and methods in sentiment analysis and challenges appear in the field.

B.Pang et al. [6] studied with the growing availability and popularity of opinion-rich resources such as online review sites and personal blogs, new opportunities and challenges arise as people now can, and do, actively use information technologies to seek out and understand the opinions of others. This survey covers techniques and approaches that promise to directly enable opinion-oriented information-seeking systems. Our focus is on methods that seek to address the new challenges raised by sentiment-aware applications, as compared to those that are already present in more traditional fact-based analysis.

Antonia Azzini in et al [2] in this research they have deliberated the repercussions of this issue for process mining methods that are based on the computation of frequency between event need. They have illustrated a procedure developing two distinct procedures. The major procedure is expected at computing the mismatch between the data sources to be assimilated. The subsequent procedure uses this information for performing a map reduce algorithm organized to integrate data in a dependable method.

Yang Song in et al [19] they have applied other use cases of the Storage Mining Framework such as storage association scheduling using correlation information between storage volumes. In accumulation, they have considering other big data machine learning platforms such as Apache Mahout [5], RHIPE [14], Ricardo [15], IBM Big-Insight with SystemML, and leveraging such platforms in other IT management aspects such as server and network management.

According to McKinsey & Co [10] Big Data is seen as the next edge for novelty, opposition and efficiency and as such the connected solicitations will contribute to financial growth. The positive influences of big data deliver a huge possibility for establishments. In directive to accomplish these objectives numerous issues should be examines and deliberated in the context of complex systems and using systems methods such as complete thinking and

system dynamics. Therefore main problems are developing and these work-in-progress efforts to discuss a few key characteristics absorbed to the growth and adopting data mining techniques and approaches for big data.

Xindong Wu in et al [18] they have respect Big Data as a developing leaning and the essential for Big Data mining is ascending in all science and engineering domains. With Big Data technologies, they have expectantly be capable to deliver most applicable and most precise social sensing feedback to improved appreciate society at real-time. They can advance motivate the contribution of the public listeners in the data production circle for societal and inexpensive events. The age of Big Data has reached.

III. PROPOSED METHODOLOGY

Big Data is a novel term used to recognize the datasets that outstanding to their large size, we cannot supervise them with the characteristic data mining software tools. Instead of defining Big Data as datasets of a concrete large size, for illustration in the organize of magnitude of petabytes, the explanation is related to the information that the dataset is too big to be manage without using new algorithms or technologies. Big data design challenges that are the important point of this work.

Reviews data is large enough so we cannot store and process them normally. So we will be using Hadoop (HDFS to store reviews data and MapReduce to Process it). Hadoop MapReduce's parallel processing capability has increased the speed of extraction and transformation of data. Hadoop MapReduce used as a data integration tool by reducing large amounts of data to its representative form which can then be stored in the data warehouse.

Our proposed analysis approach significantly out performs the cooperative update analysis in terms of the required computational complexity and communication costs, with mild compromise on the prediction accuracy. Whenever users stay on some hotels and restaurants, they provide their review based on their experience in hotel, like room quality, hotel services, food quality, value for money hotel etc. So based on that review, this approach

determines whether that review is positive or negative. We also determine ranking of individual hotels and individual rating of hotel services of each hotel. We can also use same approach for other things as well like Cars/Bikes/Laptop/Mobiles etc.

Our proposed work also provides classification of reviews data from the all reviews provided by customers. Based on positive review words and negative reviews words values decide the rating or textual review about hotels and individual all services of hotel for easily identified by customers as well as hotel manager to choose hotel as per own requirement. Store reviews data in Hadoop file system (HDFS).our proposed work provide functionality like Generate sentiment of each review, Generate overall sentiment of thing (Hotel/Restaurants/Cars/Bikes etc) based on all reviews given by all users for that thing, Generate rank for each review, Generate overall rank of the thing, Top 5 positive and Top 5 Negative reviews about the things.

Huge with heterogeneous and diverse data sources: Autonomous with distributed and decentralized control, and Complex and evolving in data and knowledge associations. To control both the computation complexity and information discussion acquire to local learners by adjusting cross correlation thresholds use to group them. Developments the problem should be framed in a more general theory. To use for illustration allowing for it from the point of observation of the conviction revision problem that revise the problem of integrate new information with previous knowledge .

Legitimate capability to appreciate not only the data structures but also the information and business value that is extracted from big data: As one of data analysis techniques, rough sets based methods have been successfully applied in data mining and knowledge discovery during last decades, and particularly useful for rule acquisition and feature selection. To our knowledge, most of the traditional algorithms based on rough sets are the sequential algorithms and corresponding tools only run on a single computer to deal with small data sets. To expand the applications of rough sets in the field of data mining and knowledge discovery from big data,

we discuss about rough set based parallel methods for knowledge acquisition in this paper. Based on MapReduce, we design corresponding parallel algorithms for knowledge acquisition on the basis of the characteristics of the data. The proposed algorithm is implemented on Hadoop platform Comprehensive experiments are conducted to evaluate the proposed algorithms and the results demonstrate that our algorithms can effectively process large scale data sets.

3.1 Proposed Model

Using HDFS and based on different configurations of Hadoop, I have proposed the work flow model for sentiment and predictive analysis of big data as shown in figure 2 below, the model for the mining of relative information from customer provided reviews big data shown in figure 3 below and Component& interface of big-data solution reference in figure 4 below.

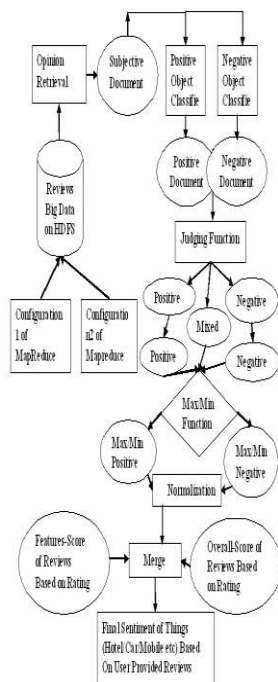


Figure 2: Proposed Model for Sentiment and Predictive analysis of Big Data

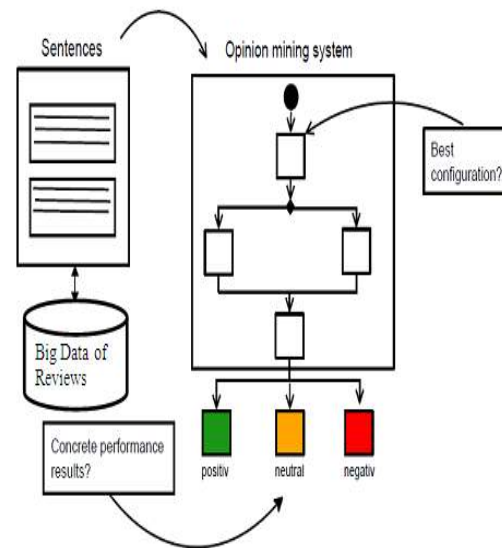


Figure 3: Mining of Relative Information from Big Data

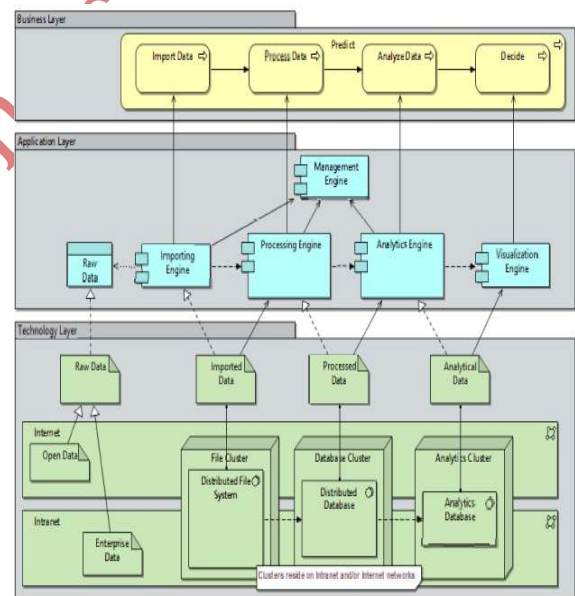


Figure 4: Component & Interface of Big-Data Solution Reference

IV. EXPERIMENTS AND RESULTS

We have obtained result by comparing approach based on different-different configuration of HDFS in hadoop by executing an application. Up to this point, several approaches have been selected from literature and are implemented. It has been tried here to mention the model approach which are

fundamentally different from each other. We built our approach with three different configuration of MapReduce in hadoop environment for comparing and analyzing result of all. While comparing found three advantages of System Based On hadoop:

- Execution time taken in config 2 & 3 of HDFS in hadoop for our proposed approach is approximately 30% less than other configuration of hadoop for map reduces.
- Execution Memory usage is 25% less in config 1 & 3 for our designing framework, so we can consider using config 3 which is good in performance and better in memory usage for sentiment and predictive analysis of big data approach.
- Proposed approach can be ported without modification on many Hadoop infrastructures.
- Proposed framework can be testing with vast amount of data up to terabytes or exabytes because this framework based on hadoop environment.

Based On Job Time Taken: Below show the graph between number of jobs and Execution time of proposed framework with Different configuration of Hadoop.

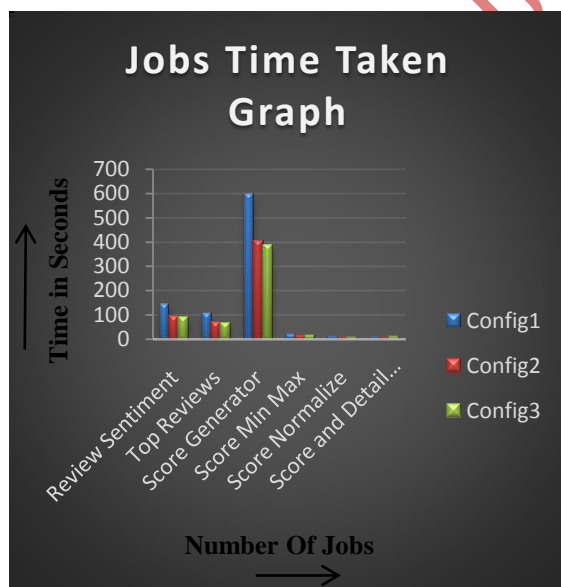


Figure 5: Job Time Taken Graph in HDFS

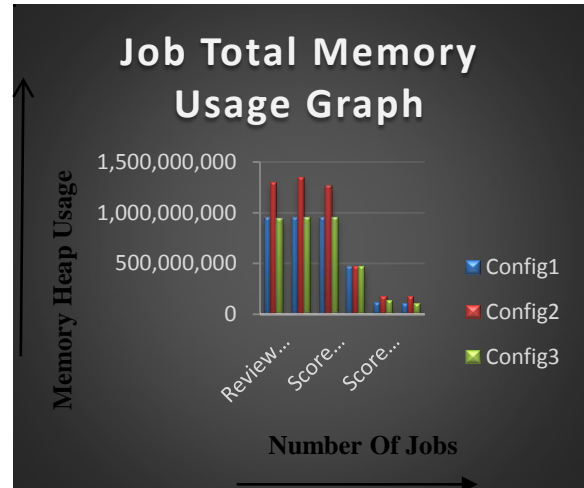


Figure 6: Job Total Memory Usage Graph in HDFS

Based On Job Total Memory Usage: Above show the graph between number of jobs and Execution memory usage of proposed framework with Different configuration of Hadoop.

Comparison Based on Running Time:

Performance of time taken in config 2 & 3 of HDFS in hadoop for our proposed approach is approximately 30% less than other configuration of hadoop for map reduces based on execution time. So we can consider using config 3 which is good in performance and better in time taken for sentiment and predictive analysis of big data approach. As shown in figure 7 below:

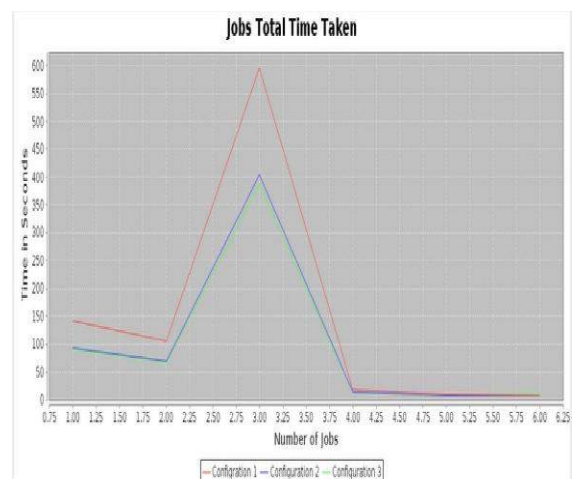


Figure 7: Comparison Graph Based On Running Time

Comparison Based on Memory:

Performance of memory usage is 25% less in config 1 & 3 for our designing framework based on execution memory. So we can consider using config 3 which is good in performance and better in memory usage for sentiment and predictive analysis of big data approach. As shown in figure 8 below:



Figure 8: Comparison Graph Based On Memory Uses

Execution time comparison for sentiment and predictive analysis of big data approach based on different configuration of HDFS in hadoop shown by Table 1:

Table 1 Execution Time Comparison Table

Job Name	Config 1		Config 2		Config 3	
	Time Taken (Sec)	Total Heap Usage(MB)	Time Taken (Sec)	Total Heap Usage(MB)	Time Taken (Sec)	Total Heap Usage(MB)
Review-Sentiment	142	903.0625	92	1231.9375	91	892
Top-Reviews	105	901.75	70	1277.375	68	902.5
Score-Generator	595	906.375	403	1198.625	386	905.875
Score-Min-Max	19	444	14	448.75	14	445.125
Score-Normalize	9	101.25	8	160.187	8	123.00781
Score-and-Detail-Merge	8	93.3125	8	158.812	9	93.375

Execution memory comparison for sentiment and predictive analysis of Big Data approach based on

different configuration of HDFS in Hadoop shown by Table 2:

Table 2 Execution Memory Comparison Table

Configuration	No of Mappers Per node	No of reducers per node	Maximum Heap per task(childopts)
Config 1	2	2	200MB
Config 2	4	2	400MB
Config 3	4	2	200MB

By designing framework following facts we have:

Execution time taken in config 2 & 3 of HDFS in hadoop for our proposed approach is approximately 30% less than other configuration of hadoop for map reduces. Execution Memory usage is 25% less in config 1 & 3 for our designing framework, so we can consider using config 3 which is good in performance and better in memory usage for sentiment and predictive analysis of big data approach.

Sentiment analysis of reviews Framework designed with Apache Hadoop scale with data and compilation to overcome the problem of data storage complexity. This Framework can be tested with vast amount of data (TB).

The Enhanced Framework used single node cluster for hadoop is efficient in programming for parallel execution of application. But we can also increase performance in multi node cluster, that's the advantage of using hadoop.

Any user will get actual rating and top review that will help him to choose the best option based on his/her requirement.

Hotels Management/Car Makers etc. can use the system to improve their service based on the ranking and top reviews we provide.

V. CONCLUSION AND FURTHER WORK

We use and apply our propose consequences identify, The problem of a customer in identifying the hotel and its services based on requirement can be reduced by proposing a new analysis approach which will be the resultant of big data; which is fast and accurate.

We have focused on the textual reviews of customers rather than scalar reviews to get the best possible option irrespective of rating of hotels in a city. The proposed framework tries to integrate both structured and unstructured data from all the reviews provided by different users. This approach not only provides an accurate consequence but also requires a relatively small computational time. Finally, it can be concluded that the performance and efficiency of sentiment analysis of big data can be increased by using Big Data and data mining computing technique. Amongst the study of various analyses, it has been found suitable, efficient and adoptable computing technique which is combination of sentiment and predictive analysis of big data. This technique is implemented by using HDFS in Hadoop. By implementing big data computing, data storage management is also significantly enhanced. At last, user provided reviews' data for different hotels and their services are analyzed by using map reducers in HDFS etc. to make access faster. We can reduce time and total memory using map reducer in HDFS so that it can be accessed faster and easily and query responses also become faster. Our proposed analysis approach can flexibly balance amongst computational complexity, communication cost and prediction accuracy based on the necessities of the application and the end users.

How to make Sentiments classification performance more domains independent?, To answer this question, further work on extending dictionary, used by Sentiments classification algorithm with lexicon words from different domains should be done. This task requires huge effort, as all ambiguous words which have a tonality for one context but are neutral in another should be eliminated.

VI. REFERENCES

- [1] V.D. Nguyen, B.Vargheseb and A.Barkerb," The Royal Birth of 2013: Analysing and Visualising Public Sentiment in the UK Using Twittera," *IEEE International Conference on Big Data*, 2013.
- [2] A. Azzini and P. Ceravolo," Consistent Process Mining Over Big Data Triple Stores," *IEEE International Congress on Big Data-978-0-7695-5006-0/13*, 2013.
- [3] A. Ghoting, R. Krishnamurthy, E. Pednault, B. Reinwald, V. Sindhvani, S. Tatikonda, Y. Tian, and S. Vaithyanathan, "SystemML: Declarative machine learning on MapReduce," *IEEE ICDE*, 2011.
- [4] A. Lazarevic and V. Kumar, Feature bagging for outlier detection, *KDD '05*, 2005.
- [5] Apache Mahout, <http://mahout.apache.org>.
- [6] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, pp. 1-135, 2008.
- [7] B. Liu, E. Blasch, Y. Chen, D. Shen and G. Chen, "Scalable Sentiment Classification for Big Data Analysis Using Naïve Bayes Classifier," *IEEE International Conference on Big Data*, 2013.
- [8] C. Perlich and G. Swirszcz, on cross-validation and stacking: building seemingly predictive models on random data, *ACM SIGKDD Explorations Newsletter*, 12(2010), pp. 11-15.
- [9] G. Vinodhini*, R. Chandrasekaran," Sentiment Analysis and Opinion Mining: A Survey," *International Journal of Advanced Research in Computer Science and Software Engineering*, (Volume 2, Issue 6), ISSN:- 2277 128X, June 2012.
- [10] McKinsey Big Data: The next frontier for innovation, competition and productivity- Global Institute (2011).
- [11] M. Sewell, Ensemble learning, UCL Research Note, 2007.
- [12] N. D. Valakunde, Dr. M. S. Patwardhan," Multi-Aspect and Multi-Class Based Document Sentiment Analysis of Educational Data Catering Accreditation Process," *International Conference on Cloud & Ubiquitous Computing & Emerging Technologies-2013 IEEE*, 2013.
- [13] NESSI (2012) Big Data – A New World of Opportunities, White Paper.
- [14] P. Buhlmann and B. Yu, Boosting with the L2 loss: Regression and classification, *J. American Stat. Assoc.*, 98(2003), pp. 324-339.
- [15] RHIPE, <http://www.datadr.org>.

[16] S. Das, Y. Sismanis, K. Beyer, R. Gemulla, P. Haas, and J. McPherson, "Ricardo: integrating R and Hadoop," ACM SIGMOD, 2010.

[17] S. McConnell and D. Skillicorn, Building predictors from vertically distributed data, CASCON '04, 2004.

[18] X. Wu1, X. Zhu, G. Wu, and W. Ding," Data Mining with Big Data" Knowledge and Data Engineering," *IEEE Transactions on* (Volume:PP , Issue: 99)- Page(s):1 ISSN :-1041-4347, 26 June 2013.

[19] Y. Song, G. Alatorre, N. Mandagere, and A. Singh," Storage Mining: Where IT Management Meets Big Data Analytics," *IEEE International Congress on Big Data*, 2013.

[20] Z. Liu, "Map Reduce-based Back propagation Neural Network over large scale mobile data," *Natural Computation (ICNC), 2010 Sixth International Conference on* (Volume: 4).

IJournals