DOI: http://dx.doi.org/10.26483/ijarcs.v8i8.4850

ISSN No. 0976-5697

Volume 8, No. 8,September-October 2017

International Journal of Advanced Research in Computer Science

CASE STUDY

Available Online at www.ijarcs.info

BIG SENTIMENT ANALYSIS USING K-MEANS CLUSTERING: A SURVEY

Shalini Yadav
Department of Computer Science and Engineering
Parul University, Gujarat, India

SunitaYadwad
Department of Computer Science and Engineering
Parul University, Gujarat, India

PrakshiYadav
Department of Electronics and Communications
LNMIIT, Jaipur, India

Abstract: With the extending areas for social events, online overviews, and long-range relational correspondence, the present work is to investigate reviews, evaluations, and trades on the Web so the customer can settle on aninformed decision. Conclusion investigation, otherwise called opinion mining is the computational investigation of sentiments, assumptions, and feelings communicated in common dialect preparing and message examination. Opinion mining, otherwise called Sentiment analysis, assumesan imperative part of this procedure. It is the investigation of feelings, i.e., Assumptions, Expressions that areexpressed in regular dialect. Normal dialect methods areconnected to separate feelings from unstructured information.There are a few procedures which can be utilized to examination such sort of information. Here, we areordering these methods extensively as ”supervised learning, ”unsupervised learning” and ”hybrid techniques.”Both learning methods are combined to get the benefits ofunstructured data in huge volumes. The goal of this paperis to give the review of Sentiment Analysis with K-Meansclustering, their difficulties and a similar examination of its methods. In this paper sentiment analysis is collaborated with parallel K-Means for processing massive amount of data and to extract benefits of parallelization.

Keywords: SVM (Support Vector Machine), Naive Bayes,K-Means, Hadoop, MapReduce.

1 INTRODUCTION

Data mining is a multidisciplinary field, drawing work from areas including machine learning, database technology, knowledge-based systems, statistics, pattern recognition, information retrieval, neural networks, artificial intelligence, high-performance computing and data visualization. Sentiment analysis is the computational investigation of individuals’ feelings, examinations, dispositions, and attitudes toward elements, people, issues, occasions, points and their properties. The investigation of aggregate conduct is to see how people act in an interpersonal interaction condition. Seas of information produced by online networking like Facebook, Twitter, and YouTube. It presents openings and difficulties to ponder aggregate conduct on an extensive scale. It expects to figure out how to foresee aggregate conduct in web-based social networking. For instance, organizations dependably need to find open or customer feelings about their items and administrations. Steps to perform sentiment analysis is shown in Fig. 1

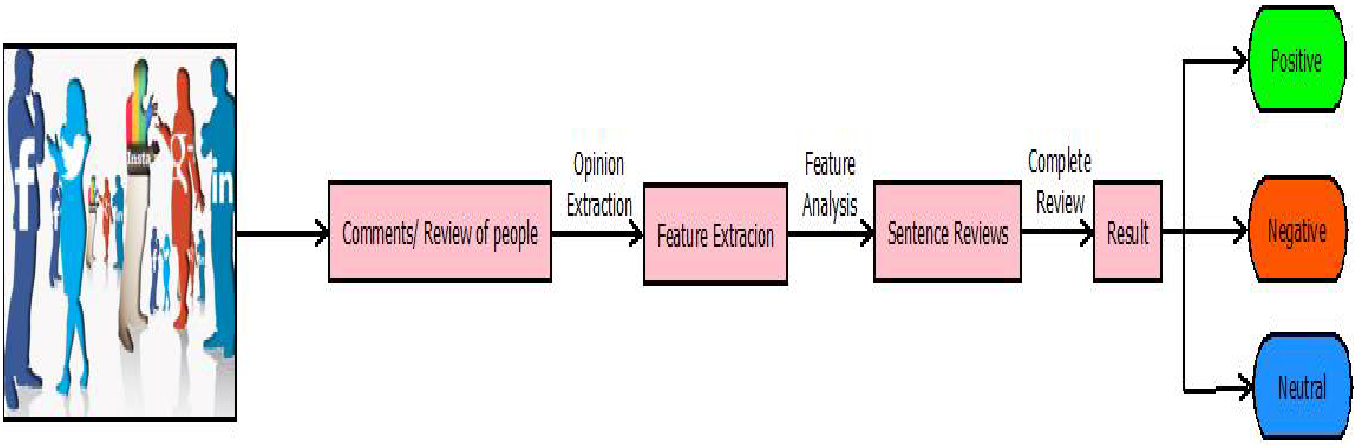

Figure 1: Process of Sentiment Analysis

Potential clients additionally need to know the assessments of existing clients previously they utilize a policy or buy an item. With the dangerous development of web-based social networking (i.e., audits, gathering exchanges, online journals and informal organizations) on the Web, people and associations are progressively utilizing general conclusions in these media for their basic leadership. Notwithstanding, finding and observing sentiment destinations on the Web and refining the data contained in them remains an imposing errand as a result of the multiplication of assorted locales. Each website normally contains an enormous volume of obstinate content that is not effortlessly deciphered in long gathering postings and sites. The normal human will have difficulty distinguishing significant destinations and precisely abridging the data and conclusions contained in them. Also, it is additionally realized that human investigation of content data is liable to significant predispositions, e.g., individuals regularly give careful consideration to feelings that are predictable with their inclinations. Individuals likewise have difficulty, attributable to their mental and physical restrictions, creating steady outcomes when the measure of data to be handled is expensive. Automized opinion mining and rundown frameworks are in this manner required, as subjective inclinations and mental constraints can be overwhelmed with a target slant examination framework. In the previous decade, a lot of research has been done in the scholarly community. There are likewise various business organizations that give feeling mining administrations.

Clustering/segmentation are a standout among essential methods utilized as a part of data mining. K means clustering bunches comparable perceptions in groups keeping in mind the end goal to have the capacity to separate bits of knowledge from huge measures of unstructured information.

When you need to examine the Facebook/Twitter/YouTube remarks of a specific occasion, it is difficult to physically take a glance at every single say and see where the conclusion concerning a specific brand/occasion/individual untruths.

Numerous unsupervised grouping calculations are produced over the time. In these all K-means is for the most part used for its straightforwardness and capability. With the brisk change of web, the data we need to process is growing in petabytes. K-Means has restricted preparing capacity in light of its chance many-sided quality in the serial situation. The two noteworthy weakness downsides of K-means are 1) It relies upon beginning group focus which was taken haphazardly. To beat the principal condition that is beginning bunch focus many changed variations of K-means have been proposed and to conquer the neighborhood ideal issue many papers with the mix of k-means with some developmental methods have been proposed [1].

And furthermore recognized that K-means isn't reasonable for preparing monstrous measure of information due to its chance many-sided quality in the serial situation. To defeat this issue, numerous specialistshave brought parallelization into the calculations. In [2] it proposes the parallel K-Means utilizing MPI (Message Passing Interface). It causes us better the time multifaceted nature of vast datasets. In [3] [4] [5] it proposes Parallel K-Means Clustering Based on Map Reduce where the proposed estimation was associated beneficially for significant datasets, and test outcomes were laid out.

The sentiment analyses utilized for a look at and gather or sort the general population feelings, audits, opinions, feelings about the item, occasion, administrations and so forth that are in human dialect. It arranges them in positive, negative or common relies upon people's slants, sentiments, feelings that communicated in it, for example, word that is sure in one circumstance could likewise be considered as antagonistic in an alternate circumstance, take a word "long." If client said that telephone life is long, that is sure feeling, or if the customer said booting time of telephone is extensive, that would be a negative supposition.

2 DATA SOURCES

This part shows brief points of interest of data sets. Sites, audit sites, information and small-scale sites give a unique making sense of to deliverable phase of the items and administrations gave to clients. Customer's feeling is an exceptional standard for the change of the Sentiment examination. Web journals, survey locales, information, and small-scale web journals give a decent comprehension of the gathering level of the items and administrations.

2.1 Blogs

The name related to weblog sites aludes as the blogosphere. With an expanding utilization of the web, blogging and blog pages are developing quickly. Blog pages have turned into the most famous intends to express one's genuine beliefs. Individuals compose concerning the topics they need to impart to others on a blog. Blog pages have ended up being the across the board way to deal with express ones' individual suppositions about any item or point.

2.2 Review sites

For the client in settling on the choice of obtaining, the suppositions of others are a vital factor. An enormous amount of purchaser produced reports is close by on the web. The commentator's information used in loads of the feeling grouping ponders gathered from the web-based business web locales like www.amazon.com

2.3 DataSet

That is the accumulation of the surveys that are utilized for the order of the feelings for instance motion picture audits informational collections that include film appraisal with the capacity of thousand +ve and one thousand - ve prepared film encounters. The informational indexes are having the different sorts of surveys of the items like camera, advanced cells, gadgets, books and so forth.

2.4 Twitter Dataset

Twitter is a social news site.The Twitter Sentiment Analysis Dataset contains characterized tweets; each column is set apart as 1 for positive opinion and 0 for negative sentiment, and twitter dataset is accessible at http://socialcomputing.asu.edu/datasets/Twitter

3 THE EXISTING TECHNIQUES FOR SENTIMENT ANALYSIS USING K-MEANS CLUSTERING

The essential thought of K Means clustering is to shape K seeds, to begin with, and after those gathering perceptions in K groups on the premise of separation with each of K seeds. The perception will be incorporated into the nth seed/group if the separation between the perception and the nth seed is least when contrasted with different seeds.

3.1 The Process of building K clusters on Social Media content information

A simple K-means clustering is shown in Fig. 2 Below is a brief overview of the methodology involved in performing a K Means Clustering Analysis.

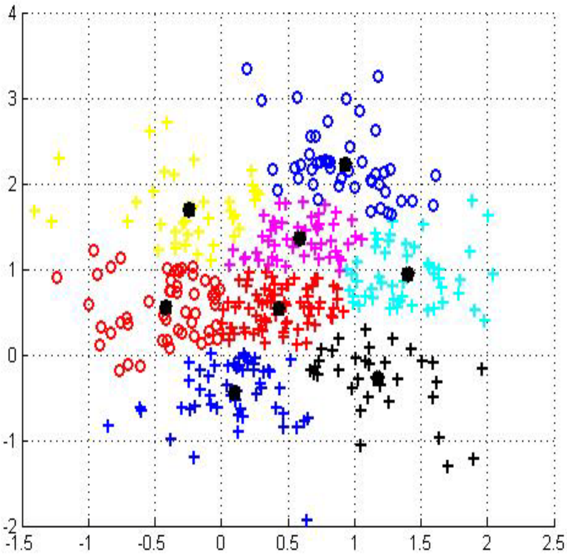

Figure 2: Simple K Means Clustering

- The initial step is to pull the online networking notices for a specific time allotment utilizing web-based social networking listening instruments (Radian 6, Sysmos, Synthesio and so on.). You would need to assemble question/add catchphrases to pull the information from online networking Listening apparatuses.
- The subsequent stage is data cleansing. This is the most critical part of web-based social networking remarks don't have a particular arrangement. Individuals utilize local people/slangs and so on via web-based networking media to express their feelings, so it's vital to have the capacity to see through them and comprehend the fundamental assessment.
- Expel punctuations, numbers, stop words (R has particular stopword library however you can likewise make your own rundown of stopwords). Additionally, expel copy columns or URLs from the online networking notices.
- The subsequent stage is to make corpus vector of the considerable number of words.
- When you have made the corpus vector of words, the following stage is to make a report term matrix.

3.2 Literature survey of papers on sentiment analysis using k-means and parallel k-means

1. In the paper [6] author has actualized and demonstrated the outcomes for three distinctive clustering calculations for estimation analysis.All three calculations give a successful time to perform assumption examination for a various and huge number of tweets. K-means calculation is performed for 2 and 3 groups. In future, it can be actualized for huge number clusters. Cure calculation additionally can be utilized for various situations for a look and perform changed bunching. Birch calculation can be utilized all the more proficiently to perform grouping and give the better outcome for slant analysis.In future conclusion, examination can be streamlined utilizing distinctive methods for handling enormous information like Hadoop, flexible inquiry and investigation.
2. As per McKinsey Report [7] Big Data can be characterized "data sets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze" .Those datasets are created for the most part through the internet utilization, cell phones, sensor systems, endeavor framework, and association. Enormous information isn't just about volume it likewise considers assortment and speed. Enormous information produced can be organized, unstructured and semi-organized.
3. With present-day innovations, for example, gadgets, hardware, vehicles installed with sensors and increment of web utilization among masses particularly "online networking, for example, Facebook and furthermore GPS gadgets" producing the expansive measure of unstructured information which is mind-boggling. This is called as Big Data [8].

The proposed strategy in this paper points how to enhance the nature of sentiment analysis on textual document surveys utilizing Hadoop structure. Additionally, the strategy depends on preparing and testing will enhance the precision of after effects of an investigation. The attention is on the utilization of open source advancements mainly [9]. Numerous clients give surveys for the single item. Such a great many surveys can be broken down and analyzed using big data. The outcomes can be introduced in an advantageous visual shape for the non-special reviews audits given for the item in the guide MapReduce system. Numerous past explores the territory the area of sentiment analysis utilized twitter information to arrange the tweets into positive or negative, depending on the supposition. Various scientists have attempted to enhance the exactness of such classification. In this paper, an approach of 'Cluster-then-Predict' is utilized to first cluster the tweets utilizing k-means calculation and then perform characterization utilizing Classification Trees. This grouping operation makes the information space particular, which brings about the formation of better prescient models which has led to a more accurate classification of sentiments [10].

4 INTRODUCTION TO BIG DATA SOLUTIONS FOR SENTIMENT ANALYSIS

The use of the web is driving to the age of huge informational indexes. Productive treatment of such vast information (otherwise called Big Data) is a continuous critical research over the world. Treatment of Big Data incorporates capacity and preparing of Big Data. Likewise examining the Big Data, this incorporates finding the example or learning disclosure from the Big Data which is called as "Big Data analysis."Increase in Internet utilization is, for the most part, a direct result of the web-based social networking fame. Different Big information advances like schema-less databases or NoSQL databases, Hadoop, Hive, Pig, PLATFORA goes for collection, processing and putting away of enormous information in a less expensive and compelling way. Huge information does not just mean about putting away the huge measure of information yet also putting away and analyzing [11].

Hadoop is an adaptable open source structure where Hadoop innovation causes us to perform operations on conveyed information in an effective way. Hadoop contains a programming model called Map Reduce where it gives a related execution to handling and producing enormous informational indexes with parallel, conveyed calculation on a group. In this paper, we are taking suppositions of the general population of a notable individual. Individuals communicated their perspectives about the individual which encourages us to break down the positive, negative and nonpartisan remarks.

ECO SYSTEMS OF HADOOP:

- PIG
- HIVE
- HBASE
- SQOOP
- FLUME
- OOZIE

1. Pig gives a motor to executing information streams in parallel on Hadoop. It incorporates a Dialect, Pig Latin, for communicating these information streams. Pig Latin incorporates administrators for a considerable lot of the customary information operations (join, sort, channel, and so on.), and additionally the capacity for clients to build up their own capacities for perusing, handling, and composing data.Pig is an Apache open source venture. This implies clients are allowed to download it as source or parallel, utilize it for themselves, add to it, and—under the terms of the Apache License—utilize it in their items and change it as they see fit.

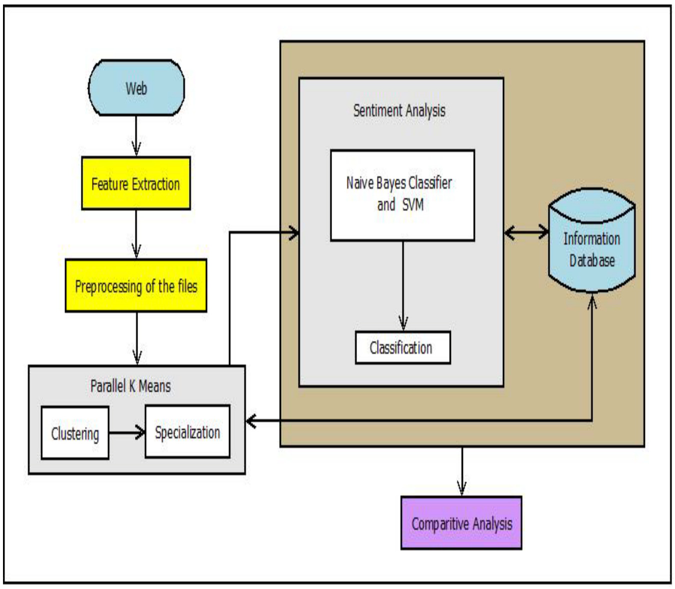

Figure 3: Proposed framework

2. Hive gives a rich arrangement of devices in different dialects to perform SQL-like information investigation on information put away in HDFS. The brilliant individuals at Facebook have contributed Hive to the Apache Project. As of the distribution of this book, Hive is experiencing dynamic improvement. Flume is a conveyed, solid, and accessible administration for effectively moving a lot of information not long after the information is delivered. This discharge gives a versatile course to move information around a group and additionally solid logging. The essential utilize case for Flume is a logging framework that accumulates an arrangement of log records on each machine in a group and totals them to a unified relentless store, for example, the Hadoop Distributed File System (HDFS).
3. Oozie is a work process scheduler framework to oversee Apache Hadoop employments. Oozie Workflow occupations are Directed Acyclical Graphs (DAGs) of activities. Oozie Coordinator employments are intermittent Oozie Workflow occupations activated by time (recurrence) and information accessibility

The objective of this paper is to give a prologue to this interesting issue and to show a structure which performs notion examination by partner adjusted K-Means calculation in parallel utilizing MapReduce with Naïve Bayes arrangement and Support vector machine.

Fig. 3.demonstrates the proposed system of our approach which has four modules: Data extraction, Pre-preparing, Clustering, and Classification. In this proposed approach, a number of steps are utilized conceptualize, plan and play out a viable opinion investigation of online cell phone audits that will be accomplished by partner changed K-Means calculation with Naïve Bayes Classification and Support Vector Machine and to assess which strategy is more exact for slant examination.

The technique of conclusion examination has following advances:

- First step: First concentrate the information to be investigated from the web. In our work, we have extricated information from twitter cell phone database.
- Second step: For preprocessing and extremity count of the information we have made a preparation dataset for positive, negative, normal assessment words and stopped words in SQL.
- Third step: Preprocessing-In the pre-preparing of the information the words which are not conveying any assessments or conclusion are expelled from the information. Another assignment performed in pre-handling is stemming. It is the way toward diminishing determined words into their root frames, e.g., word satisfaction is decreased into root frame upbeat. Along these lines, after pre-handling, we get just the important information on which we can without much of a stretch apply the systems.
- Fourth Step: "Ascertain Polarity" gives us the tally of positive, negative and normal conclusion words in the

entered information which is utilized by the systems as a contribution for additionally preparing.

- Fifth Step: Clustering is the way toward shaping the groups of the information, questions inside a bunch have comparable properties, however, they are not like protests in different groups. By analyzing the notable k-Means calculation parallel k-Means needs one sort of MapReduce work. The undertaking of doling out each specimen to the nearest centroid should be possible utilizing the guide work and computing the new centroid position should be possible utilizing lessen work. In k-Means the most tedious part is removed computation part. We have to compute separate between one example and all centroids. On the off chance that the given dataset has N-measurement tests and the coveted bunch number is K, it will require K*N separate figuring in every cycle. So with the most extreme cycle number I, the aggregate estimations expected to finish the bunching undertaking will be K*N*I. So we have to influence this assignment of calculation to parallel for this first we have to store the dataset in stable stockpiling, i.e., HDFS. So that the dataset is part of a number of parts and all-inclusive communicated to all mappers. Subsequently, remove figuring can be executed parallel in a number of machines. Lessen capacity will figure the mean of all specimens and gives another centroid position. The nitty-gritty clarification of Map and Reduce capacities given beneath.

  Map task takes two sections an info one is the dataset put away in HDFS and Take k center points from the dataset. To begin with, the dataset split into some parts each split will send to the mapper as a <*Key, Value*> sets.At each mapper, it calculates the distance between one sample and all the centers and then assigns a sample to the closest cluster according to the distance. The middle results are a pair of <*Key,Value*> sets where a key is the id of the nearest center and value is the sample. To enhance the performance of the reducer parceled by the key on the mapper and afterward sent to the reducer.

  The Input of the reducer is the middle of the result which is created by the mappers. The key, Value sets which are having a similar key will send to a similar reducer. The key is the id of the cluster and value is the list of the sampler assigned to the cluster. At that point reducer will aggregate every one of the samplers in the group and gather the samplers which are having a similar key. At that point, by figuring the mean, the new centroids will be created. At that point, the new centroids are sent for the following emphasis.

- Sixth step: Classification is finished utilizing Naïve Bayes Classification and Support Vector Machine. Naive Bayes Classification depends on administered learning. It is a factual technique for an arrangement. It processes the probabilities of the results to decide if an example has a place with a specific class or not. It is utilized for both symptomatic and prescient issues. Support Vector Machine depends on administered learning. It has related learning calculations which are utilized for performing assignments, for example, information examinations, design acknowledgment and is utilized for grouping and relapse investigation. SVM display speaks to cases as focuses in space, mapped so the cases of the distinctive classifications are characterized by a wide hole which is as vast as could be expected under the circumstances.

Initial a survey is entered and afterward pre-preparing of the information is done as such as to expel all the good for nothing information. It helps continuously feeling examination by decreasing the clamor of the information and enhances the classifier execution and increment the speed of the arrangement procedure. In the second step include extraction utilizing parallel K-Means is done to make classifiers working compelling; the measure of information to be examined is lessened, and pertinent highlights are known which are valuable in the order process. In the last stage grouping systems have been connected Naive Bayes Classification and Support Vector Machine. Both the systems gives result as a likelihood work

Since the different grouping methods developed over the time have neglected to demonstrate their faultless productivity and streamlined outcome in the field, it may be that parallel programming strategy like MapReduce holds the answer for this issue also. We utilize the upside of combining K-means with and parallel programming strategies to build up another way to deal with give great quality groups.

MapReduce is a programming model and a related execution for handling and creating substantial datasets that is agreeable to an expansive assortment of genuine assignments. Clients determine the calculation as far as a guide and a decreased work, and the hidden runtime framework consequently parallelizes the calculation crosswise over huge scale groups of machines, handles machine disappointments, and calendars between machine correspondence to make proficient utilization of the system and plates.

First, a review is entered and then pre-processing of the data is done to remove all the meaningless data. It helps in real time sentiment analysis by reducing the noise of the data and improves the classifier performance and increase the speed of the classification process. In the second step feature extraction using parallel K-means is done to make classifiers working effectively; an amount of data to be investigated is reduced as well as relevant features are known which are useful in the classification process. And furthermore recognized that K-means isn't appropriate for preparing gigantic measure of information given its opportunity unpredictability in the serial situation. To beat this issue, numerous specialist has brought Parallelization into the calculations. In [4] it proposes the parallel K-Means utilizing MPI (Message passing Interface). It encourages us better the time unpredictability of expansive datasets. In [5] it proposes Parallel K-Means Clustering Based on MapReduce where the proposed figuring was associated profitably for generous datasets and test outcomes were delineated. As needs are the time the many-sided quality of K-means diminished basically and memory need of a solitary machine is evacuated because it dispersed over the gathering of machines. In [4] [5] they utilized MapReduce to parallelize the K-means Algorithm It shows that the strategy is of high versatility and productivity. In the last phase classification techniques have been applied Naïve Bayes Classification and Support Vector Machine. Both the techniques give result in the form of a probability function

5 CONCLUSION

The primary goal behind our work is investigation of various opinion mining strategies that can be utilized in sentiment analysis using K-Means clustering. A few procedures of opinion mining classifiers have been attempted from time godlike. Maybe a couple is characterized in this work which has risen as of late for productive and successful. By applying opinion mining strategies in us can distinguish the issue and propose restorative treatment. The paper is to assess the execution of the characterization calculations using K-Means clustering. Since we there are limitations in simple K-Means Clustering and majority of the data produced by the social media comprises of information that is semi-organized, unstructured and organized, we proposed a simple K-Means clustering by using parallel Map Reduce as we can load more data in parallel. This method meets the prerequisite of huge information. Consequently the paper proposes conveying of the previously mentioned expectation strategies in parallel utilizing Map Reduce.

6. REFERENCES

[1] Weizhong Zhao1, 2, Huifang Ma1, 2, and Qing He1, “Parallel K-Means Clustering Based on MapReduce” Springer-Verlag Berlin Heidelberg 2009.

[2] Fahim Ahmed M, “Parallel Implementation of K-Means on Multi-Core Processors” GESJ: Computer Science and Telecommunications 2014.

[3] Jeffrey Dean and Sanjay Ghemawat, “MapReduce: Simplified Data Processing on Large Clusters” Google white paper to appear in OSDI 2004.

[4] P K Pedireddla, SA Yadwad “An Effective and Efficient Clustering Based on K-Means Using MapReduce and TLBO”. Proceedings of the Second International Conference on Computer and Communication Technologies,2016,Pages 619-628 Publisher Springer India.

[5] Pavan Kumar, Mummareddy and Suresh Chandra Satapathy, “An Hybrid Approach for Data Clustering Using K Means and Teaching Learning Based Optimization” Springer International Publishing Switzerland 2015 S.C. Satapathy et al. (eds.), Emerging ICT for Bridging the Future, Volume 2,165 Advances in Intelligent Systems and Computing 338, DOI: 10.1007/978-3-319-13731-5 19.

[6] ShivaniRana,“SENTIMENT ANALYSIS BY ASSOCIATING MODIFIED K MEANS WITH NAÏVE BAYES CLASSIFICATION AND SUPPORT VECTOR MACHINE",International Journal of Advanced Research in Science and Engineering.

[7] Hsinchun Chen, Roger H. L. Chiang, Veda C. Storey, "BUSINESS INTELLIGENCE AND ANALYTICS: FROM BIG DATA TO BIG IMPACT ," MIS Quarterly Vol. 36 No. 4, pp. 1165-1188/December 2012.

[8] Sun, N., et al. "iCARE: A framework for big data-based banking customer analytics." IBM Journal of Research and Development 58.5/6 (2014): 4-1.

[9] Concepts and Methods of Sentiment Analysis on Big Data M. Edison 1, A. Aloysius 2 International Journal of Innovative Research in Science,Engineering and Technology (An ISO 3297: 2007 Certified Organization)Vol. 5, Issue 9, September 2016.

[10] Rishabh Soni1, K. James Mathai2,"Effective Sentiment Analysis of a Launched Product using Clustering and Decision Trees", International Journal of Innovative Research in Computer and Communication Engineering (A High Impact Factor, Monthly, Peer Reviewed Journal) Vol. 4, Issue 1, January 2016.

[11] MugdhaJinturkar , PradnyaGotmare,"Sentiment Analysis of Customer Review Data using Big Data: A Survey" International Journal of Computer Applications (0975 – 8887) Emerging Trends In Computing 2016