Sentiment Analysis of Customer Product Reviews Using Hadoop and Machine Learning Algorithms

Dissertation submitted in fulfilment of the requirements for the Degree of

MASTER OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

By

MIR AAMIR HAMID

11607487

Supervisor

MR. MAMOON RASHID

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

November 2017

TOPIC APPROVAL PERFORMA

School of Computer Science and Engineering

Program : P172::M.Tech. (Computer Science and Engineering) [Full Time]

COURSE CODE : CSE548 **REGULAR/BACKLOG :** Regular **GROUP NUMBER :** CSERGD0339

Supervisor Name : Mamoon Rashid **UID :** 20574 **Designation :** Assistant Professor
Qualification : _____ **Research Experience :** _____

SR.NO.	NAME OF STUDENT	REGISTRATION NO	BATCH	SECTION	CONTACT NUMBER
1	Mir Aamir Hamid	11607487	2016	K1637	9872245141

SPECIALIZATION AREA : System Architecture and Design **Supervisor Signature:** _____

PROPOSED TOPIC : Sentiment Analysis of Customer Product Reviews Using Hadoop and Machine Learning Algorithms

Qualitative Assessment of Proposed Topic by PAC		
Sr.No.	**Parameter**	**Rating (out of 10)**
1	Project Novelty: Potential of the project to create new knowledge	7.00
2	Project Feasibility: Project can be timely carried out in-house with low-cost and available resources in the University by the students.	7.33
3	Project Academic Inputs: Project topic is relevant and makes extensive use of academic inputs in UG program and serves as a culminating effort for core study area of the degree program.	6.83
4	Project Supervision: Project supervisor's is technically competent to guide students, resolve any issues, and impart necessary skills.	7.67
5	Social Applicability: Project work intends to solve a practical problem.	6.83
6	Future Scope: Project has potential to become basis of future research work, publication or patent.	7.17

PAC Committee Members		
PAC Member 1 Name: Gaurav Pushkarna	UID: 11057	Recommended (Y/N): Yes
PAC Member 2 Name: Er.Dalwinder Singh	UID: 11265	Recommended (Y/N): Yes
PAC Member 3 Name: Harwant Singh Arri	UID: 12975	Recommended (Y/N): Yes
PAC Member 4 Name: Balraj Singh	UID: 13075	Recommended (Y/N): Yes
PAC Member 5 Name: Raj Karan Singh	UID: 14307	Recommended (Y/N): NA
PAC Member 6 Name: Harleen Kaur	UID: 14508	Recommended (Y/N): NA

PAC Member 7 Name: Sawal Tandon	UID: 14770	Recommended (Y/N): NA
PAC Member 8 Name: Tejinder Thind	UID: 15312	Recommended (Y/N): Yes
DAA Nominee Name: Kuldeep Kumar Kushwaha	UID: 17118	Recommended (Y/N): Yes

Final Topic Approved by PAC: **Sentiment Analysis of Customer Product Reviews Using Hadoop and Machine Learning Algorithms**

Overall Remarks: Approved

PAC CHAIRPERSON Name: 11024::Amandeep Nagpal **Approval Date:** 04 Nov 2017

11/29/2017 1:50:45 PM

DECLARATION STATEMENT

I hereby declare that the research work reported in the dissertation proposal " Sentiment Analysis of Customer Product Reviews Using Hadoop and Machine Learning Algorithms" in partial fulfilment of the requirement for the award of Degree for Master of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab is an authentic work carried out under supervision of my research supervisor Mr. Mamoon Rashid. I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University's Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

MIR AAMIR HAMID

REG.NO: 11607487

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the M.Tech Dissertation proposal "**Sentiment Analysis of Customer Product Reviews Using Hadoop and Machine Learning Algorithms"**, submitted by **MIR AAMIR HAMID** at **Lovely Professional University, Phagwara, India** is a bonafide record of his original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

Mr. Mamoon Rashid

Date: 30-11-2017

Counter Signed by:

1) **Concerned HOD:**
   HoD's Signature: _____

   HoD Name: _____

   Date: _____

2) **Internal Examiner**

   Signature: _____

   Name: _____

   Date: _____

ACKNOWLEDGEMENT

To start with, I thank Almighty Allah for showering His unlimited blessings on us, for giving us the power to think and work. I would like to express my sincere gratitude and appreciation to all those who gave me the possibility to complete this Report (fulfillment of course requirement of M.Tech CSE). My Respectful Thanks goes to Mr. Mamoon Rashid, my mentor and guide, for his unparalleled support and concern and for helping me to gain confidence in myself. I would like to thank my parents, who have always been there for me with all their love and care that is worth millions. I would like to express an earnest gratitude to the faculty of Department of Computer Science & Engineering, for their helping hand time to time. Last but not the least I am extremely grateful to my friends for their support and encouragement throughout.

Mir Aamir Hamid

Table of Contents

LIST OF FIGURES

ABSTRACT

A lot of data is generated from multiple sources. This data contains many hidden patterns and information. Data from Social Networks mostly contains opinions of general public. This data is very useful for organisations. These opinions can be mined from this data. One approach is to use Sentiment Analysis. Data can be dumped into HDFS of hadoop and then classification algorithms are applied to decode the sentiment in this data. Naïve Bayes and Decision Tree Algorithms are used in this research. Naïve Bayes is a powerful and simple classification algorithm. But it assumes independence of features. So, Decision Tree can be used in conjunction with it to get more accurate results.

CHAPTER 1

INTRODUCTION

As of late, there has been a significant increase in growth of data. Data is produced from multiple sources like automobiles, banks, sensors, day-to-day human activities, social media etc. But the volume of data has grown beyond the computing power of traditional approaches of processing. There is the concept of Big Data. Big Data has the main characteristics of volume, velocity and variety. It also encompasses variability, value, veracity and volatility which help in determining the importance of big data. Big data can be either in structured or unstructured form. It can be either in a pre-stored form or can be generated in real-time. The attributes like value, volatility and veracity help in determining the importance of the data and usually data is important as it may give us useful insights regarding what decisions should a company make and what market strategy will be best. It can even allow political parties to target masses based on issues. It helps an organisation in predicting the sales behaviour. But older approaches of processing are unable to unravel the full potential of this data. They find just the portion of what actually is hidden inside the data. So, there has been an increase in approaches as to how Big Data can be processed. One such approach is the MapReduce. Hadoop is one way to implement MapReduce. Hadoop finds usage in industrial as well as academic research purposes.

Data is generated in large quantities on social media. Previously, information used to spread into little circles. Nowadays, people use social media to express their opinions about everything. They post about things and share images. Social network have received an upward surge and data generated from them is attaining higher values day by day. According to a survey, in one minute lakhs of tweets are sent, thousands of images are shared on Facebook and lakhs of videos are watched on YouTube [3].

Twitter is one of the most used social network. It allows users to tweet using a limited number of words so that it is read by everybody. There are tweets regarding many things including business establishments, movies, political parties, educational institutions, scientific projects etc. These

tweets reflect sentiment of the people regarding various topics, products, movies. Every person has his own opinion regarding a topic or product. This sentiment gets reflected when the person tweets about that topic. The person may point out what were the positive things and what were negative ones. Companies will benefit immensely by getting this sentiment data as they can do target marketing and improve their market survey and research. There is a proposition that public opinion affects the stock price [4]. Market indicates the stock price, and what a person feels about anything will affect the market. So the sentiment of general public towards the company will impact the stock price of the company.

The existing tools are unable to handle such enormous quantities of data. The existing machines and algorithms are not able to provide the computing resources required. So, there is a need to design a model which will implement a method to capture the sentiment of people using social networks by utilising social data in conjunction with Big Data.

One of the approaches is to do Sentiment Analysis. We get data from a social network and store it into Hadoop. Then we implement a classification algorithm using MapReduce framework which will classify the data we have gathered based on what sentiments are hidden in it.

1.1 How can data be classified

Text can be classified according to what information is present in the text. Text classification can be done in two phases. In first phase, a model is built by training it with data we know what class it belongs to. First, we need to generate the dataset along with labels, then this data needs to be processed. Then model is trained after the data is vectorised. In the second phase, we test the model by making it to classify previously unseen data. This data also needs to be processed and vectored before feeding it to the model. The various basic steps included are:

- First the data is collected and is assigned to classes. Class has its own label. Then this data is split into two halves, one for each phase.
- Then this data is changed like stop words, punctuation marks etc. are removed.
- After that, the data needs to be converted in a form understandable by the computer. This is called Vectorization. Actually, some features need to be selected from the data and then data is represented using those selected features.

- Then we need to select an algorithm which will classify our data and then we feed one half of data to the algorithm. The model learns from this data in this phase. This is the first phase.
- Then we feed the next half of data to the model and let the model classify it.

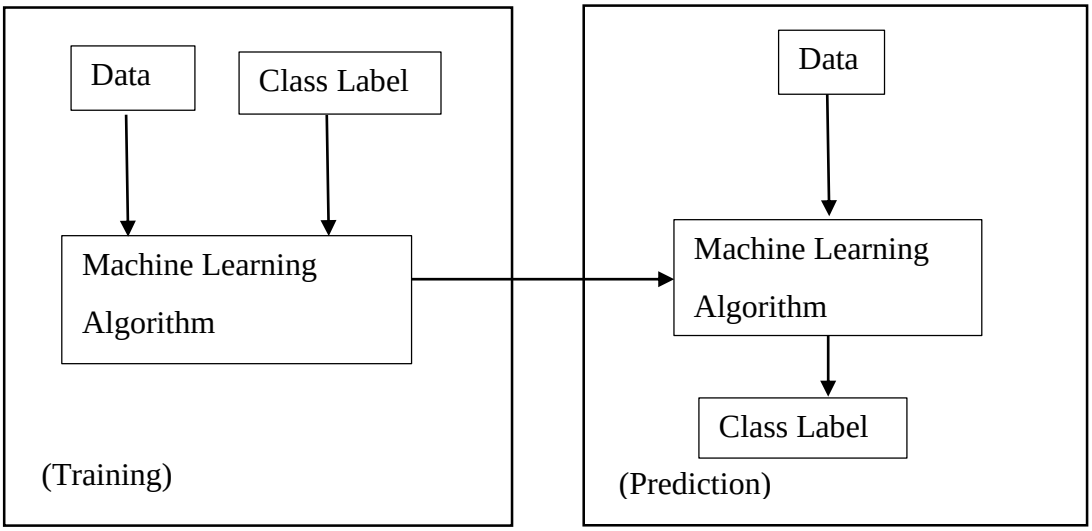

Figure 1.1: Basic idea of how data is classified

1.2 Definition and Need for Sentiment Analysis

Sentiment analysis means identifying the opinions of a person about anything. Sentiment analysis is of use in social media monitoring as we can gain an understanding of what the general public feels about something. Sentiment analysis can be used in various ways. Organizations use this approach to gain insights from social data. Same can be said about stock market. It is believed that the Obama administration used sentiment analysis to tailor announcements and design appropriate campaign messages in 2012 presidential election. There are numerous methods to check what the opinion of the person towards a topic is. One approach is to make two groups of data, positive sentiment and negative sentiment, and assign all the data to one of the two classes. This can be done using lexicon that does its job according to what words are present in the data. Here, all the

data is traversed through and those words that correspond to some sentiment are found. There is a predefined dictionary which contains whether the word is reflecting good sentiment or bad sentiment and each data item is given a rating also. Each and every part of data undergoes this to get the overall rating and sentiment. But, to get overall sentiment based on predefined dictionary is not feasible as there will be difficulty in keeping up the predefined dictionary. This led to the development of classification algorithms which are employed to classify data [20], which include Decision Table, Naive Bayes, Maximum Entropy, Decision Tree, K-Means and SVM. All of these require data with known class labels. This data will be input to the model for training. The model learns from this data. It will generate a model according to this information. The model can then be able to give a predicted value when new information is then given to it.

1.2.1 Feature Vector

For classifying the data, certain features of the data need to be selected. This is called feature vector or term vector. It is done so that computer can understand the data. This needs to be done during training the model as well as during testing it. We need to convert all the data from the tweets into feature vectors.

1.2.2 Some Classification Algorithms

- **Naive Bayes**

To group data into classes, one classification algorithm is the Naïve Bayes Algorithm [16].
In naïve bayes, we will have the number of groups. Also we need to choose a number of data points irrespective of their class. We need to determine prior probability of every group. This depends upon the data in the class and also the whole data. Prior probability defines the probability of a data item to belong to this group. Then we use the previously chosen number of data points and use it when a new data item comes. For each data item, we have to compute the likelihood. This determines the probability of the new data item to belong to either of the groups. It is also computed for new data item corresponding to each group. Then from the prior probability and the likelihood, we determine the Posterior Probability of data item corresponding to each class. The highest value indicates that the new data item belongs to that class.
Basically, it is based on Bayesian Theorem of conditional probability. But it assumes that features are independent. So, it is called naïve.

- **SVM**

SVM tries to determine a line in the plane of the data which will be able to isolate the data of different groups. There can be a linear line, or plane or even a hyperplane which will do this isolation. The separator is selected based on distance. The separator with largest normal distance from the data is chosen.

- **Decision Table**

Decision table is a compact way to represent information in a concise manner and predict class labels for the same. It is particularly useful when there are large number of features considered. Many features are simultaneously considered and a label is predicted. It becomes easy when there are a lot of features and when we need to remove some combinations of features.

- **Decision Tree**

Decision tree [5] a tree-like structure and an alternative to decision tables which maps data items to their predicted class labels. It is composed of two types of nodes: Decision nodes that check the value of attributes and leaf nodes that show the predicted label. It consists of a series of branches corresponding to yes or no for each selected attribute. We choose a starting node and check the condition there relative to the value. This allows us to select a particular branch. It allows us to move from start node to a decision node. Then we check the conditions at the decision node and select the appropriate branch. In this manner we follow a path from start node to the leaf node which shows the class label.

However, decision trees become too complex when there are many features. Also, when there need to be some features removed, we need to prune the tree. Hence decision table is a good alternative.

1.3 Hadoop Cluster

Hadoop is a free programming framework. It is a part of the apache project of the Apache Software Foundation. It is a java-based framework which is capable of processing very huge datasets. Hadoop runs in a distributed computing environment and it can process unstructured, heterogeneous data coming at high volume.

1.3.1 Overview

A hadoop cluster is composed of racks. The number of nodes in one rack is not fixed and varies from 20 to 40. These racks are connected by a single network connecting device. There are master nodes and slave nodes in the hadoop cluster. Mater node is associated with managing the Hadoop Distributed File System. It also has the knowledge about which slave nodes contain the data of a particular file. The slave node is associated with carrying out the task. When a job is to be performed, the slave nodes are tasked with retrieval and processing of the data When a task begins, the slave node will find the location of data from the master node and retrieve the data and then process it.

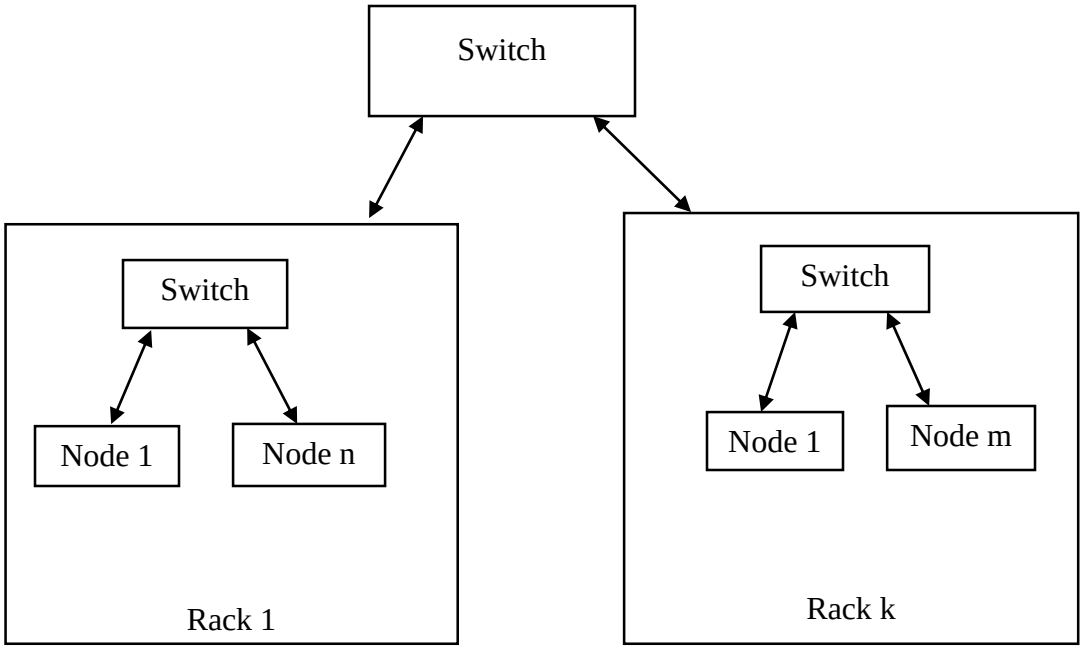

Figure 1.2: Overview of a Hadoop Cluster

1.3.2 Hadoop Distributed File System (HDFS)

HDFS (Hadoop Distributed File System) is the filesystem for the hadoop framework. It is distributed and scalable. It consists of Name Node and Data Node. Name node is responsible for management of the filesystem and data node is responsible for storage and retrieval of data.

Also, HDFS stores the data in terms of blocks that have a fixed size and stores them all over the cluster. Since Hadoop utilises commodity hardware, there can be failures. So, the data is stored on multiple nodes redundantly to ensure high availability. By default, each data item will be stored thrice [4]. Hadoop uses rack awareness to store the replicas. Rack Awareness means that the closest data node is chosen by the name node. A record of rack IDs of datanodes is kept by namenode and that helps in determining the nearest datanode. One replica is stored on the local rack. The second replica is stored on another datanode. The third replica is stored in a different rack.

The nodes operate in a master slave relationship. The NameNode reflects the master and keeps track of all the DataNodes where the blocks for a given file are located. Data nodes also pass the information of the data blocks to the NameNode.

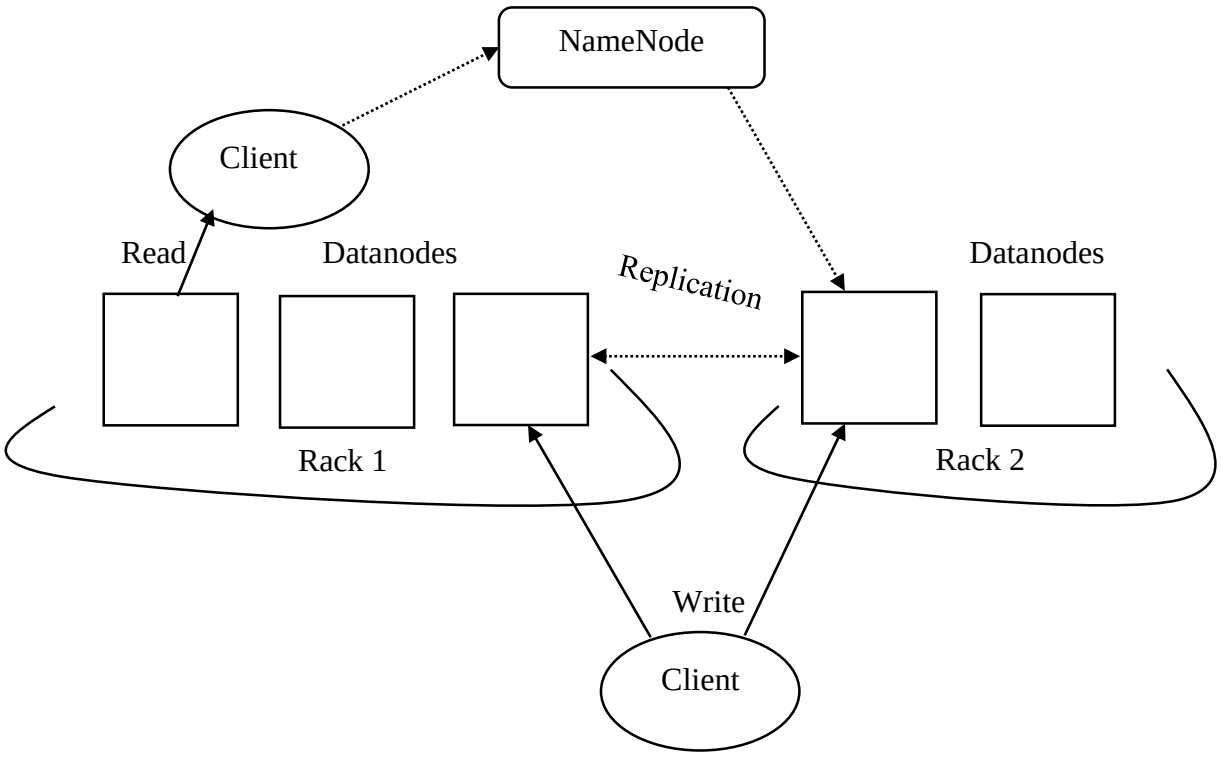

Figure 1.3: HDFS Overview

Since there is replication in HDFS, it provides some benefits. It provides high availability. If one node fails, Hadoop continues to perform the job by shifting to the replica node. It also helps in cluster-rebalancing by rebalancing nodes which have high demand. Also, HDFS provides fault

tolerance due to its ability to detect failed nodes. All data nodes send a signal to the main node. When any data node doesn't send the signal, the main node marks the data node as failed.

1.3.3 MapReduce

MapReduce is a software framework which is associated with processing of data through a cluster of nodes. There are two types of nodes in MapReduce- JobTracker and TaskTracker. They also run on master-slave model. JobTracker is the master node and TaskTracker is the slave node. There are two functions in MapReduce; Map function and Reduce function. The Map divides a large job across the worker nodes of a cluster. Multiple map functions can be executed at once, so it's the part of the program that divides up tasks. The reduce function then takes the output of the map functions, and does some processing on them to generate the desired result. This final result is the answer to the original query. Every map and reduce functions are independent of each other. All of the processing is occurring on the separate nodes where the data is located. When a job needs to be executed, it is taken by the JobTracker. The JobTracker then assigns tasks of the job to the TaskTrackers. The concept of data locality is followed here.

Data Locality means the closeness of the data to the job working on the data. Data locality has its own importance because in its absence there would be network congestion. Data Locality has three options [23]:

- Data Local: If the task is running on the same node which contains the data, it is referred to as Data Local.
- Rack Local: If the task runs on different node but in the same rack which has the data, it is called the Rack Local. Here data needs to be copied between nodes but within same rack.
- Different Rack: When the task and data are on different racks. Here data needs to be copied between racks.

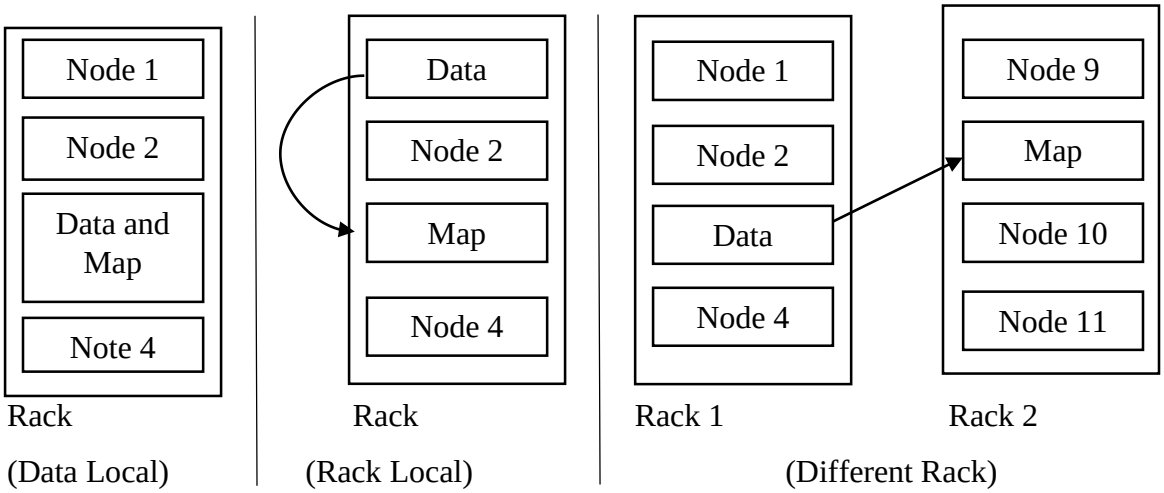

Figure 1.4: Data Locality Scenarios in MapReduce

Also, in case of failure of any TaskTracker, the task needs to be reassigned to another TaskTracker. This is done by the JobTracker. This provides fault-tolerance. All data nodes have a TaskTracker running on them. TaskTracker receives a signal from the JobTracker to specify that it is running and ready to receive tasks.

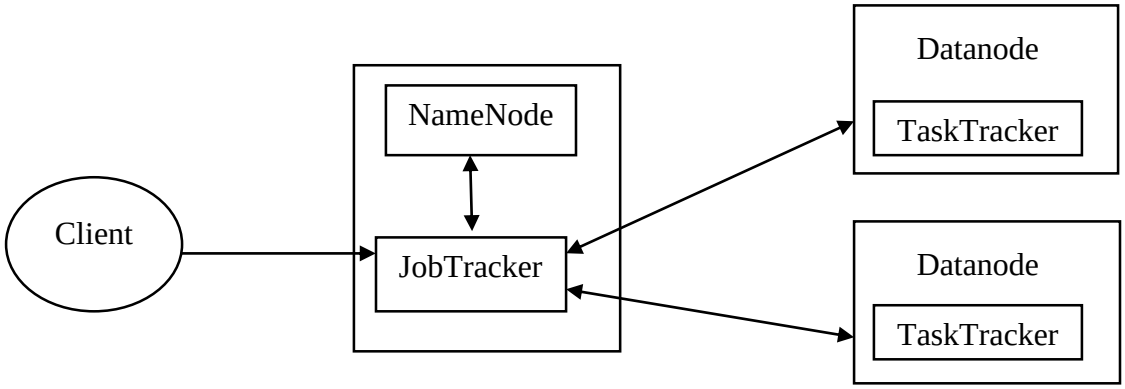

Figure 1.5: Components in a Hadoop Cluster

The following steps are performed in a Hadoop Cluster:

- JobTracker receives a MapReduce job started by the client.
- NameNode will provide the information about where the data is stored to the JobTracker.
- TaskTrackers are allocated the tasks. Here the concept of Data Locality is used.

- When the tasks are allocated to TaskTrackers, map function is implemented. The information obtained during this phase acts as input to the reduction phase. The output of reduce function is stored in HDFS.

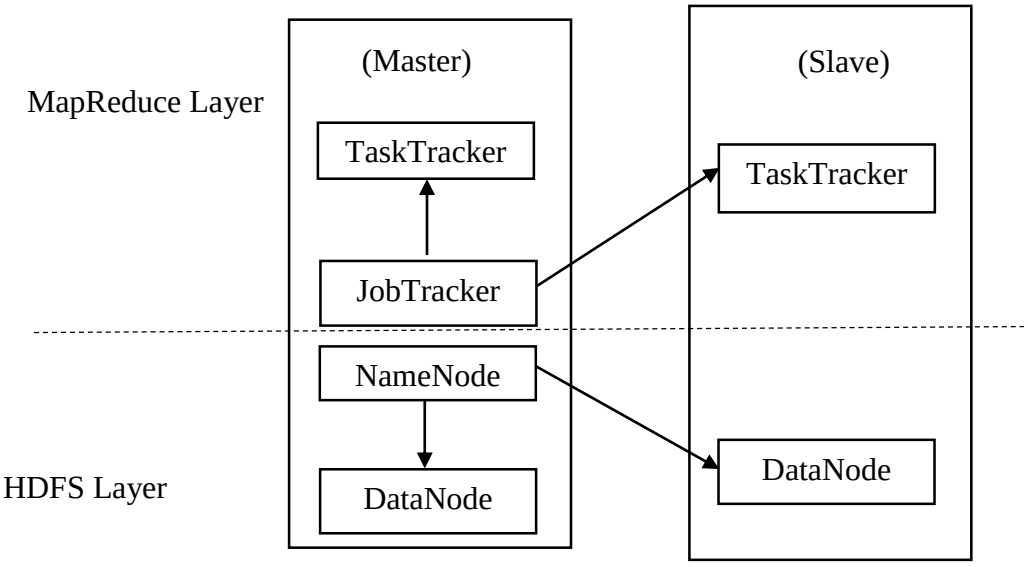

Figure 1.6: Multinode Cluster Overview of MapReduce

CHAPTER 2

REVIEW OF LITERATURE

According to Ajinkya Ingle et.al. user opinions on social networks can be used to make predictions. A cluster of nodes in hadoop will process twitter data. Analysis consists of many steps like Tokenisation, Normalisation, and Classification. Various types of dictionaries are used to determine polarity of user opinions expressed on twitter. A segment classifier will classify the tweets as positive, negative or neutral based upon the tokens in it.

According to Huma Pandey et.al. a twitter API is used to fetch data from twitter. Tweets are pre-processed like usernames, urls, special symbols, hashtags are removed and emoticons are converted into words. Then Naïve-Bayes algorithm is applied for classification, which uses a wordnet dictionary. In Map phase, polarity of tokens is generated to check overall polarity of a tweet. In reduce phase, we are categorising the polarity into specific class. Converting emoticons into equivalent words increases the efficiency of the system.

According to Divya Sehgal et.al. twitter data contains many sentiments which can be analysed using Hadoop. Twitter's API is used to get data from twitter. Then the data undergoes some processing. First removal of stop words is done. Then the tokens are changed into a structured form as they are mostly in unstructured form. Then emoticons are also translated for higher accuracy. Then Map-Reduce is used to find sentiment of each word and the sum gives the overall sentiment of the tweet.

According to Jalpa Mehta et al. Two map-reduces are used. In first mapreduce, a sentence is detected and stopwords, hashtags etc. are removed. Then we search for words which represent features and are then clustered. Then OpenNLP is used for POS Tagging . Phrase removal is done before stopword removal. In second mapreduce, a sentiwordnet dictionary is used, scores are given to words, which is averaged then to get overall value.

According to Dr. U Ravi Babu reviews given by customers on e-shopping websites have a lot of sentiment which need to be analysed. A dataset is collected and it undergoes some processing like Tokenization, Translating slang words, stemming etc. POS tagging is used in conjunction with Sentiword dictionary to calculate the sentiment.

Vaishali Sarathy et.al have discussed about a general sentiment analysis framework. It consists of steps like extracting relevant data from the source, pre-processing the data e.g. tokenisation and stemming, training the classifier, then using it to determine polarity of new data e.g. tweets and at last checking the accuracy.

Rajni Singh et.al. used Weka, on open source tool for data mining, to perform sentiment analysis for movie reviews. Data was taken from twitter and other online review platforms like IMDB. Then data needs to be pre-processed. Then naive bayes classifier is used. Then the accuracy needs to be checked.

Akshay Amolik et.al. have used the method of feature vectors to do sentiment analysis of twitter movie reviews. First a dataset needs to be created using tweets about movies. Then after pre-processing, feature vectors are created. Here first features of twitter are extracted, then the tweet is represented into keywords. Then a classification algorithm is used like Naive Bayes or support vector machine. Support Vector Machines have high accuracy than Naïve Bayes.

Abhinandan P Shirahatti et.al. have proposed a system of using twitter data for sentiment analysis using Hadoop. Flume is used to get data from twitter after creating a twitter application. Then it is queried using Hive.

According to Hardi Rajnikant Thakor, various classification algorithms can be used for sentiment analysis. Decision trees have fast fitting speed and fast prediction speed, but have low accuracy. Naïve bayes has high accuracy but has slow prediction speeds and consumes much time in training.

Priya. V et.al has examined the sentiment of youngsters regarding the floods in Chennai in 2016. They used flume to get data from twitter and applying Naïve Bayes algorithm. They developed a dictionary to compare the tweets with and get a sentiment score.

Mrigank Mridul et.al. compared the time taken by a file to process in Java with the time taken in Hadoop. They compared files of different sizes, with results showing that Hadoop is the fastest approach.

Bingwei Liu et.al. showed that due to growth in dataset, there is no decrease in accuracy of Naïve Bayes. Naïve bayes is able to scale up. Also, the number of cases which are classified

correctly and the number of cases which are classified wrong decrease when the size of dataset grows. Also when the dataset increases, it benefits from parallelisation of Hadoop, so processing time is reduced.

Shivangi Sharma combined Naïve Bayes with decision tables for data mining. The results show that this combination resulted in improved statistics and various parameters were valued higher in combination than in Naïve Bayes. Also, it classified more correctly, resulting in more accurate information.

M. Edison et.al showed various methods and concepts of sentiment analysis on big data. It has two approaches- Lexicon based and Machine Learning based. The machine learning approach is more popular and uses various supervised and unsupervised learning algorithms.

CHAPTER 3

PROBLEM DEFINITION

Content development in the Internet lately has made an enormous volume of data accessible. This data is exhibited in various formats, for example, posts, news articles, remarks, and surveys. Particularly in the automobile, hardware and film segments, clients have composed audits about items or their features. By gathering and examining these surveys, new clients discover others' opinion about various highlights of the item. They can contrast the items with each other to locate the best one that addresses their issues. Additionally, makers will discover qualities and shortcomings of their items or those of their contenders. In this way, makers will tackle the reported issues and utilize the business knowledge behind the examination for future speculations. The Naive Bayes algorithm has been implemented based on MapReduce framework to check sentiment and different metrics have been checked using data from Twitter. But there are some shortcomings in Naïve bayes. Its assumption that every attribute is independent leads to degradation of results and accuracy is decreased. Also, it will predict the value based on the probability only; some decision rules can be combined with the probability in order to make more correct predictions.

CHAPTER 4

SCOPE OF THE STUDY

The scope of the study of this technique is to combine naïve bayes and decision table to implement hybrid classification algorithm. To implement this technique various sentiment analysis techniques in big data are analysed and tweets data is collected and pre-processed form the twitter API. The collected dataset from Twitter is in the form of raw data that needs to be filtered in order to do classification on the data. For the Filtration of the raw data; various URL's, hashtags, punctuation marks, stop words etc. needs to be removed. Classification is done using Map Reduce platform. There is a requirement of changing the algorithms so that they are appropriate for the MapReduce model. In the proposed hybridized algorithm comprising of naïve bayes and the decision table; Decision table will store the conditional probabilities for the naïve bayes algorithm. Attributes are chosen and based on them, information is stored in the Decision table. During prediction making, decision table will make use of this model. There is a probability for every entry in the table.

CHAPTER 5

OBJECTIVES OF THE STUDY

1. To store and organize unstructured datasets from twitter into HDFS of Hadoop.
2. To design and implement hybrid classification algorithm comprising of Naïve bayes based on probability and Decision Table based on decision rules.
3. To analyze and compare the performance of the proposed approach based on accuracy with the existing naïve bayes.

CHAPTER 6

PROPOSED RESEARCH METHODOLOGY

1. **Collection of data:** Data can be collected from Twitter Data source using Twitter API. Twitter has more than 200 million month to month dynamic clients which results in billions of tweets every week. One more imperative cause of using tweet data is that tweets are mostly in text, while on others, there are usually images, videos etc.

2. **Pre-Processing and Filtering:** The collected dataset from Twitter is in the form of raw data that needs to be filtered in order to do classification on the data. For the Filtration of the raw data; various URL's, hashtags, punctuation marks, stop words and Digital words needs to be removed.

3. **Classification using Map Reduce Platform:** There is a requirement of changing the algorithms so that they are appropriate for the MapReduce model. In the proposed hybridized algorithm comprising of naïve bayes and the decision table; Decision table will store the conditional probabilities for the naïve bayes algorithm. Decision table will store the conditional probabilities for the naïve bayes algorithm. Attributes are chosen and based on them, information is stored in the Decision table. During prediction making, decision table will make use of this model. There is a probability for every entry in the table. Naïve Bayes makes use of conditional probability and bayes theorem and it assumes that features are independent. The overall class probability is estimated by combining the estimated probability.The goal of the proposed approach is to construct the model which will predict labels, so only training data needs to be tested. It will require two jobs, with both having a map phase and a reduce phase. In the first job, we will estimate the probability of the data item that it belongs to either of the classes. In the second phase, we will estimate the probability that a random data item will be allocated to the class.

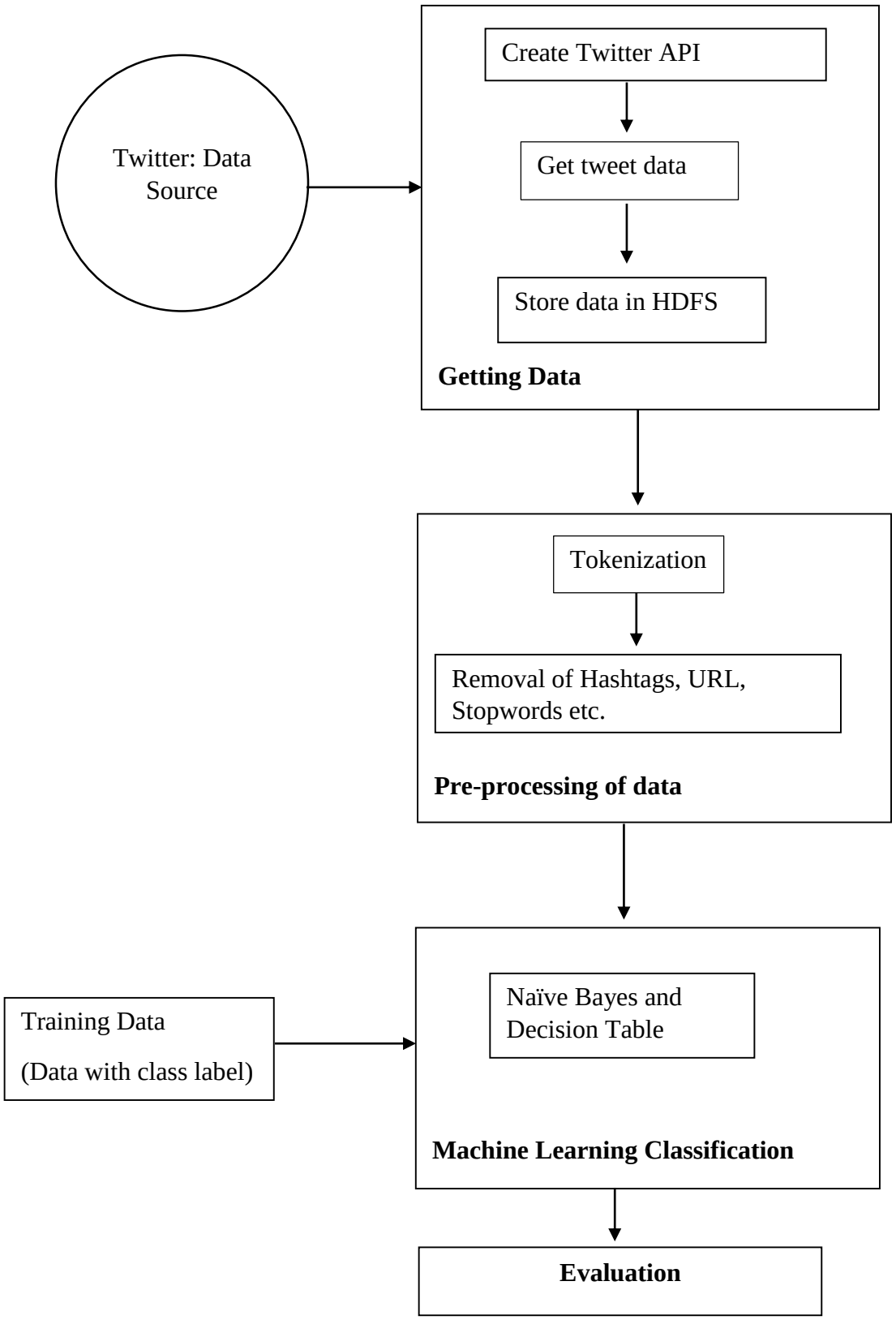

Figure 6.1: Flow chart of the proposed methodology

CHAPTER 7

EXPECTED OUTCOMES

Comments, critiques and opinion of the people play an important role to determine whether a given population is glad with the product, offerings. It allows in predicting the sentiment of a wide kind of people on a selected event of interest just like the review of a movie, their opinion on numerous topic roaming around the arena. These data are essential for sentiment analysis. In order to discover the overall sentiment of populace, retrieval of statistics from assets like Twitter, Facebook, Blogs are critical. Sentiment Analysis utilising Naive Bayes has been implemented based on MapReduce and various parameters have been assessed. But the Naïve bayes algorithm assumes that all the attributes or features are conditionally independent when class label is given; due to which the results degrade. Also, it will predict the value based on the probability only; some decision rules can be combined with the probability in order to make more correct predictions. The expected outcomes of this study is to the compare the performance of hybrid classification algorithm comprising of naïve bayes and decision table based on accuracy with the existing naïve bayes algorithms.

CHAPTER 8

CONCLUSION

Data is produced from multiple sources and it is ever-increasing. This has given rise to the concept of Big Data. Big Data has the main characteristics of volume, velocity and variety. Social network have received an upward surge and data generated from them is attaining higher values day by day. Twitter is one of the most used social network. There are tweets regarding many things including business establishments, movies, political parties, educational institutions, scientific projects etc. These tweets reflect sentiment of the people regarding various topics. Companies will benefit immensely by getting this sentiment data. So we need to classify the data based on sentiment. One classification algorithm is the Naïve Bayes Algorithm. But it assumes that features are independent. Some decision rules can be combined with the probability in order to make more correct predictions. Hadoop is a free java-based programming framework capable of processing very huge datasets. One of the approaches is to do Sentiment Analysis. We get data from a social network and store it into Hadoop. Then we implement a classification algorithm using MapReduce framework which will classify the data we have gathered based on what sentiments are hidden in it.

REFERENCES

[1] Manyika, James, et al. "Big data: The next frontier for innovation, competition, and productivity." (2011).

[2] Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters." Communications of the ACM 51.1 (2008): 107-113.

[3] Temple, Krystal. "What Happens in an Internet Minute?." Inside Scoop (2012).

[4] Shvachko, Konstantin, et al. "The hadoop distributed file system." Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on. IEEE, 2010.

[5] Yousef, Ahmed Hassan, Walaa Medhat, and Hoda Korashy Mohamed. "Sentiment Analysis Algorithms and Applications: A Survey." (2014).

[6] Dobra, Alin. "Decision Tree Classification." Encyclopedia of Database Systems. Springer US, 2009. 765-769.

[7] Ajinkiya Ingle, Anjali Kante Shriya Samak and Anita Kumari, "Sentiment Analysis of Twitter Data using Hadoop", International Journal of Engineering Research and General Science, Nov-Dec 2015, Volume 3, Issue 6, pp. 144–147.

[8] Huma Pandey and Shikha Pandey, "Sentiment Analysis on Twitter Data-set using Naive Bayes Algorithm", IEEE, 2nd International Conference on Applied and Theoretical Computing and Communication Technology, 2016, pp. 416-419.

[9] Divya Sehgal and Dr. Ambuj Kumar Agarwal, "Sentiment Analysis of Big Data Applications using Twitter Data with the Help of HADOOP Framework", IEEE, 5th International Conference on System Modelling & Advancement in Research Trends, 25th -27th November, 2016, pp. 251-255.

[10] Jalpa Mehta, Jayesh Patil, Rutesh Patil, Mansi Somani and Sheel Varma, "Sentiment Analysis on Product Reviews using Hadoop", International Journal of Computer Applications Volume 142 – No.11, May 2016, pp. 38-41.

[11] Dr. U Ravi Babu, "Sentiment Analysis of reviews for E-Shopping Websites", International Journal of Engineering and Computer Science, Volume 6 Issue 1 Jan. 2017, pp. 19965-19968.

[12] Vaishali Sarathy, Srinidhi S, and Karthika S, "Sentiment Analysis Using Big Data From Social Media", 23rd IRF International Conference, 5th April 2015, pp. 40-45.

[13] Rajni Singh and Rajdeep Kaur, "Sentiment Analysis on Social Media and Online Review", International Journal of Computer Applications, July 2015, Volume 121, Issue 20, pp. 44-48.

[14] Akshay Amolik, Niketan Jivane, Mahavir Bhandari and Dr. M Venkatesan, "Twitter Sentiment Analysis of Movie Reviews using Machine Learning Techniques", International Journal of Engineering and Technology, January 2016, Volume 7, Issue 6, pp. 2038–2044.

[15] Abhinandan P Shirahatti, Neha Patil, Durgappa Kubasad and Arif Mujawar, "Sentiment Analysis on Twitter Data using Hadoop", International Journal of Emerging Technology in Computer Science and Electronics, April 2015, Volume 14, Issue 2, pp. 831–837.

[16] Hardi Rajnikant Thakor, "A Survey Paper on Classification Algorithms in Big Data", International Journal of Research Culture Society, Volume 1, Issue 3, May 2017, pp. 21 -27.

[17] Priya. V, S Divya Vandana, "Chennai Rains Sentiment-An Analysis Of Opinion About Youngsters Reflected In Tweets Using Hadoop", International Journal of Pharmacy & Technology, Sep-2016, Vol. 8, Issue No.3, pp. 16172-16180.

[18] Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta, Kumar N, "Analysis of Bigdata using Apache Hadoop and Map Reduce", International Journal of Advanced Research in

Computer Science and Software Engineering, Volume 4, Issue 5, May 2014, pp 555-560.

[19] Bingwei Liu, Erik Blasch, Yu Chen, Dan Shen and Genshe Chen, "Scalable Sentiment Classification for Big Data Analysis Using Naive Bayes Classifier", 2013 IEEE International Conference on Big Data, pp. 99-104

[20] Shivangi Sharma, "Design and Implementation of Improved Naive Bayes Algorithm for Sentiment Analysis on Movies Review", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, Issue 1, January 2017, pp.285-291

[21] M. Edison, A. Aloysius, "Concepts and Methods of Sentiment Analysis on Big Data", International Journal of Innovative Research in Science Engineering and Technology, Vol. 5, Issue 9, September 2016, pp. 16288-16296.

[22] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002.

[23] hadoopinrealworld.com. Available Online http://hadoopinrealworld.com/data-locality-in-hadoop.