ISSN 2349-6193
Impact Factor: 3.866

IJESMR

International Journal OF Engineering Sciences & Management Research

SENTIMENT ANALYSIS USING ONLINE PRODUCT REVIEWS

R. Manivannan*[1], Dr. R. Saminathan[2] & Dr. P. Anbalagan[3]

*[1]Research Scholar, Department of Computer Science & Engineering, Annamalai University, Annamalainagar – 608 002, Tamil Nadu, India

[2&3]Assistant Professor, Department of Computer Science & Engineering, Annamalai University, Annamalainagar – 608 002, Tamil Nadu, India

Keywords: *API, SVM, OVA, GFS, HDFS.*

ABSTRACT

Sentiment analysis gained the close attention in the recent years and it is the main task of Natural Language Processing (NLP). The aim of this paper focuses to handle the problem of sentiment polarity categorization, which is the basic problem of Sentiment Analysis. We have tried to propose a general process for sentiment polarity categorization and support vector machine is used for classifying the polarity of sentences. Reviews collected from online product reviews collected from Amazon.com are the data used. The outcomes are promising for both sentence-level categorization and review-level categorization.

INTRODUCTION

Sentiment is mainly a judgment prompted by feeling. It is an attitude or thought, hence the other name opinion mining. It gives the exact idea of people's sentiment towards certain entities (products). The Internet serves as a resource to provide information regarding sentiment, people are provided with opportunities to post their own ideas, thoughts via social media, such as forums, micro-blogs, or online social networking sites. Researchers and developers release their Application Programming Interfaces (APIs) to promote data collection. For example, the Twitter has three different versions of APIs, status and user information are gathered through REST API, and the Search API helps the developers to query specific content whereas the Streaming API helps to collect Twitter content in real time. We can mix those APIs to create their own applications. Thus sentiment analysis is considered highly reliable; the availability of online data makes it possible.

On the other hand there are several difficulties like opinions differ in different situations and people do not always express opinions in a same way. We have to be very careful as the meanings change completely with respect to text, for example useful and not useful changes the complete picture. People may differ in their opinions; mostly there will be equal number of positive and negative comments in such case it is very difficult for a computer to judge. Sometime vague expressions make it even more difficult to conclude.

The users hunger is on for and dependence upon online advice and recommendations the data reveals is merely one reason behind the emerge of interest in new systems that deal directly with opinions as a first-class object. Sentiment analysis concentrates on attitudes, whereas traditional text mining focuses on the analysis of facts. There are few main fields of research predominate in Sentiment analysis: sentiment classification, feature based Sentiment classification and opinion summarization. Sentiment classification deals with classifying entire documents according to the opinions towards certain objects. Feature-based Sentiment classification on the other hand considers the opinions on features of certain objects. Opinion summarization task is different from traditional text summarization because only the features of the product are mined on which the customers have expressed their opinions. Opinion summarization does not summarize the reviews by selecting a subset or rewrite some of the original sentences from the reviews to capture the main points as in the classic text summarization.

In online review data have several flaws that potentially hinder the process of sentiment analysis. The first flaw is that since people can freely post their own content, the quality of their opinions cannot be guaranteed. For example, instead of sharing topic-related opinions, online spammers post spam on the forums. Some spam is meaningless at all, while others have irrelevant opinions also known as fake opinions. The second flaw is that ground truth of such online data is not always available. Hence there is a need to obtain like a tag of a certain opinion, indicating whether the opinion is positive, negative, or neutral [8-11].

In existing methods for sentiment analysis is categorization of sentiment polarity. Given a piece of written text, the problem is to categorize the text into one specific sentiment polarity, positive or negative (or neutral). Based on the scope of the text, there are three levels of sentiment polarity categorization, namely the document level, the sentence level, and the entity and aspect level . The document level concerns whether a document, as a whole, expresses negative or positive sentiment, while the sentence level deals with each sentence's sentiment categorization; The entity and aspect level then targets on what exactly people like or dislike from their opinions [5]. Since reviews of much work on sentiment analysis have already been included in this section upon which our research is essentially based a list of positive words and a list of negative words, respectively, based on customer reviews [13].

The proposed method consists of two important factors, namely the feature vector extraction [12], Sentiment polarity categorization based on the sentence level and review level. Parts of speech are a model which aims to classify roles according to parts of speech has been explored. In this model, information is used as part of a feature set which leads to sentiment classification on a dataset. The model parts of speech are supposed to be the significant indicator of sentiment expression and which works on subjectivity detection that represents the close relationship between presence of adjectives and sentence subjectivity. But, many experimental results show that using only adjectives as features leads to worse performance. The process of sentiment polarity categorization is twofold: sentence-level categorization and review-level categorization. Given a sentence, the goal of sentence-level categorization is to classify it as positive or negative in terms of the sentiment that it conveys. Training data for this categorization process require ground truth tags, indicating the positiveness or negativeness of a given sentence.

However, ground truth tagging becomes a really challenging problem, due to the amount of data that we have. Since manually tagging each sentence is infeasible, a machine tagging approach is then adopted as a solution. The approach implements a bag-of-word model that simply counts the appearance of positive or negative (word) tokens for every sentence. If there are more positive tokens than negative ones, the sentence will be tagged as positive, and vice versa. This approach is similar to the one used for tagging the Sentiment 140 Tweet Corpus. Training data for review-level categorization already have ground truth tags, which are the star-scaled ratings [14].

LITERATURE SURVEY

Apart from classification of positive and negative sentiments, researchers also studied the problem of predicting the rating scores (e.g., 1–5 stars) of reviews [1]. In this case, the problem can be formulated as a regression problem since the rating scores are ordinal, although not all researchers solved the problem using regression techniques. Kim, Soo-Min, and Eduard Hovy (2004) [1] experimented with SVM regression, SVM multiclass classification using the one-vs-all (OVA) strategy, and a meta-learning method called metric labeling. It was shown that OVA based classification is significantly poorer than the other two approaches, which performed similarly. This is understandable as the numerical ratings are not categorical values. Liu, Bing, Minqing Hu, and Junsheng Cheng 92005) [2] improved this approach by modeling rating prediction as a graph-based semi-supervised learning problem, which used both labeled (with ratings) and unlabeled (without ratings) reviews.

Pak, Alexander, and Patrick Paroubek (2010) [3] introduced a bag-of-opinions representation of documents to capture the strength of n-grams with opinions, which is different from the traditional bag-of-words representation. Each of the opinions is a triple, a sentiment word, a modifier, and a negator. For example, in "not very good", "good" is the sentiment word, "very" is the modifier and "not" is the negator. For sentiment classification of two classes (positive and negative), the opinion modifier is not crucial but for rating prediction, it is very important and so is the impact of negation. A constrained ridge regression method was developed to learn the sentiment score or strength of each opinion from domain-independent corpora (of multiple domains) of rated reviews. The key idea of learning was to exploit an available opinion lexicon and the review ratings. To transfer the regression model to a newly given domain-dependent application, the algorithm derives a set of statistics over the opinion scores and then uses them as additional features together with the standard unigrams for rating prediction.

Prior to this work, (Choi, Yejin, and Claire Cardie, 2009) [4] proposed an approach to extracting adverb-adjective-noun phrases (e.g., "very nice car") based on the clause structure obtained by parsing sentences into a hierarchical representation. They assigned sentiment scores based on a heuristic method which computes the contribution of adjectives, adverbials and negations to the sentiment degree based on the ratings of reviews where these words occurred. Unlike the above work, there was no learning involved in this work.

Instead of predicting the rating of each review, Sarvabhotla, Kiran, Prasad Pingali, and Vasudeva Varma (2011) studied the problem of predicting the rating for each aspect. A simple approach to this task would be to use a standard regression or classification technique. However, this approach does not exploit the dependencies between users' judgments across different aspects. Knowledge of these dependencies is useful for accurate prediction. Thus, this paper proposed two models, aspect model (which works on individual aspects) and agreement model (which models the rating agreement among aspects). Both models were combined in learning. The features used for training were lexical features such as unigram and bigrams from each review. Chesley, P., Vincent, B., Xu, L., & Srihari, R. K (2010) [6] used a similar approach as that in (Pang and Lee, 2005) [1] but with a Baysian network classifier for rating prediction of each aspect in a review. For good accuracy, instead of predicting for every review, they focused on predicting only aspect ratings for a selected subset of reviews which comprehensively evaluates the aspects. Clearly, the estimations from these reviews should be more accurate than for those of other reviews because these other reviews do not have sufficient information. The review selection method used an information measure based on Kolmogorov complexity. The aspect rating prediction for the selected reviews used machine learning. The features for training were only from those aspects related sentences. The aspect extraction was done in a similar way to that in (Bhatt, A., Patel, A., Chheda, H., & Gawande, 2015) [7].

BIG DATA AND HADOOP

Big data means really a big data; it is a collection of large datasets that cannot be processed using traditional computing techniques. Big data is not merely a data; rather it has become a complete subject, which involves various tools, techniques and frameworks. Big data involves the data produced by different devices and applications. Big data technologies are important in providing more accurate analysis, which may lead to more concrete decision-making resulting in greater operational efficiencies, cost reductions, and reduced risks for the business. To hardness the power of big data, you would require an infrastructure that can manage and process huge volumes of structured and unstructured data in real-time and can protect data privacy and security. In this approach, an enterprise will have a computer to store and process big data. In Fig.1 data will be stored in an RDBMS like Oracle Database, MS SQL Server or DB2 and sophisticated software's can be written to interact with the database, process the required data and present it to the users for analysis purpose.

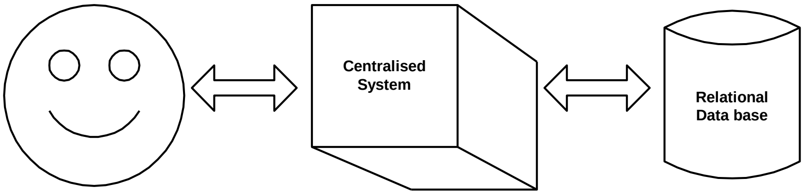


Fig. 1 Data Storage in RDBMS

This approach works well where less volume of data that can be accommodated by standard database servers, or up to the limit of the processor which is processing the data. But when it comes to dealing with huge amounts of data, it is really a tedious task to process such data through a traditional database server. Google solved this problem using an algorithm called MapReduce. This algorithm divides the task into small parts and assigns those parts to many computers connected over the network, and collects the results to form the final result dataset as shown in the Fig.2.

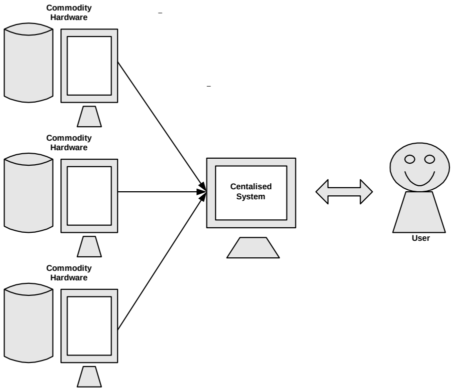


Fig. 2 Stored data in Commodity Hardware

Hadoop:

Doug Cutting, Mike Cafarella and team took the solution provided by Google and started an Open Source Project called HADOOP in 2005 and Doug named it after his son's toy elephant. Now Apache Hadoop is a registered trademark of the Apache Software Foundation. Hadoop runs applications using the MapReduce algorithm, where the data is processed in parallel on different CPU nodes [16]. In short, Hadoop framework is capable enough to develop applications capable of running on clusters of computers and they could perform complete statistical analysis for huge amounts of data as shown in Fig.3.

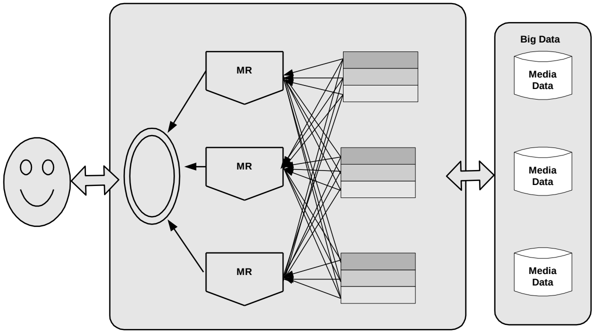


Fig. 3 Process of MapReduce

Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models [15]. A Hadoop frame-worked application works in an environment that provides distributed storage and computation across clusters of computers. Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage.

Hadoop Architecture:

Hadoop framework includes following four modules namely Hadoop Common, Hadoop YARN, Hadoop Distributed File System (HDFS) and Hadoop MapReduce as shown in Fig. 4.

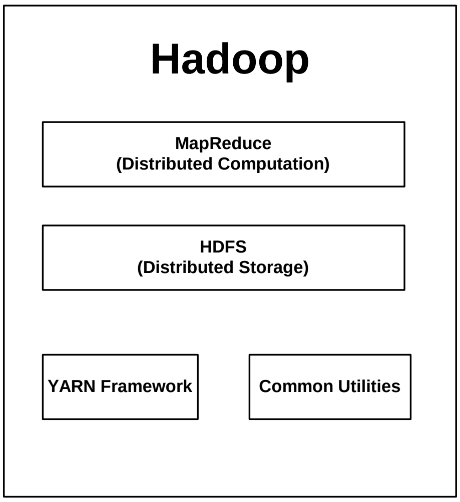

Hadoop
MapReduce (Distributed Computation)
HDFS (Distributed Storage)
YARN Framework
Common Utilities

Fig. 4 Distributed Computation and Storage with Cluster Management Technology (YARN)

Hadoop MapReduce

Hadoop MapReduce is a software framework for easily writing applications which process big amounts of data in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner. The framework takes care of scheduling tasks, monitoring them and re-executes the failed tasks. The MapReduce framework consists of a single master JobTracker and one slave TaskTracker per cluster-node. The master is responsible for resource management, tracking resource consumption/availability and scheduling the jobs component tasks on the slaves, monitoring them and re-executing the failed tasks. The slaves TaskTracker execute the tasks as directed by the master and provide task-status information to the master periodically. The JobTracker is a single point of failure for the Hadoop MapReduce service which means if JobTracker goes down, all running jobs are halted.

Hadoop Distributed File System

Hadoop can work directly with any mountable distributed file system such as Local FS, HFTP FS, S3, FS, and others, but the most common file system used by Hadoop is the Hadoop Distributed File System (HDFS). The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) and provides a distributed file system that is designed to run on large clusters (thousands of computers) of small computer machines in a reliable, fault-tolerant manner. HDFS uses a master/slave architecture where master consists of a singleNameNode that manages the file system metadata and one or more slaveDataNodes that store the actual data. A file in an HDFS namespace is split into several blocks and those blocks are stored in a set of DataNodes. The NameNode determines the mapping of blocks to the DataNodes. The DataNodes takes care of read and write operation with the file system. They also take care of block creation, deletion and replication based on instruction given by NameNode. HDFS provides a shell like any other file system and a list of commands are available to interact with the file system. These shell commands will be covered in a separate chapter along with appropriate examples.

Hadoop YARN

This is a framework for job scheduling and cluster resource management.

Hadoop Common Utilities

These are Java libraries and utilities required by other Hadoop modules. These libraries provide file system and OS level abstractions and contains the necessary Java files and scripts required to start Hadoop.

PROPOSED WORK

In the proposed method instead of removing objective content, in our study, all subjective content was extracted for future analysis. The subjective content consists of all sentiment sentences. A sentiment sentence is the one that contains, at least, one positive or negative word. All of the sentences were firstly tokenized into separated English words. Every word of a sentence has its syntactic role that defines how the word is used. The syntactic roles are also known as the parts of speech. There are 8 parts of speech in English: the verb, the noun, the pronoun, the adjective, the adverb, the preposition, the conjunction, and the interjection. In natural language processing, Part Of Speech (POS) taggers have been developed to classify words based on their parts of speech. For sentiment analysis, a POS tagger is very useful because of the following two reasons, nouns and pronouns usually do not contain any sentiment, A POS tagger can also be used to distinguish words that can be used in different parts of speech. The overall architecture of the proposed method is represented in Fig. 5.

Text Preprocessing:

Preprocessing of data is the process of preparing and cleaning the data of dataset for classification. Here is the hypothesis of having the data properly pre-processed: to reduce the noise in the text should help improve the performance of the classifier and speed up the classification process, thus aiding in real time sentiment analysis.

1. **Tokenization:** Given input as character sequence, tokenization is a task of chopping it up into pieces called tokens and at the same time removing certain characters such as punctuation marks. A token is an instance of sequence of characters that are grouped together as a useful semantic unit for processing.
2. **Stop Words Removal:** A stop-list is the name commonly given to a set or list of stop words. It is typically language specific, although it may contain words. A search engine or other natural language processing system may contain a variety of stop-lists, one per language, or it may contain a single stop-list that is multilingual. Some of the more frequently used stop words for English include "a", "of", "the", "I", "it", "you", and ”and” these are generally regarded as 'functional words' which do not carry meaning. When assessing the contents of natural language, the meaning can be conveyed more clearly by ignoring the functional words .Hence it is practical to remove those words which appear too often that support no information for the task.
3. **Stemming:** It is the process for reducing derived words to their stem, or root form. Stemming programs are commonly referred to as stemmers or stemming algorithms. A simple stemmer looks up the inflected form in a lookup table, this kind of approach is simple and fast. The disadvantage is that all inflected forms must be explicitly listed in table.eg. “developed”, “development” , ”developing” are reduced to the stem “develop”.

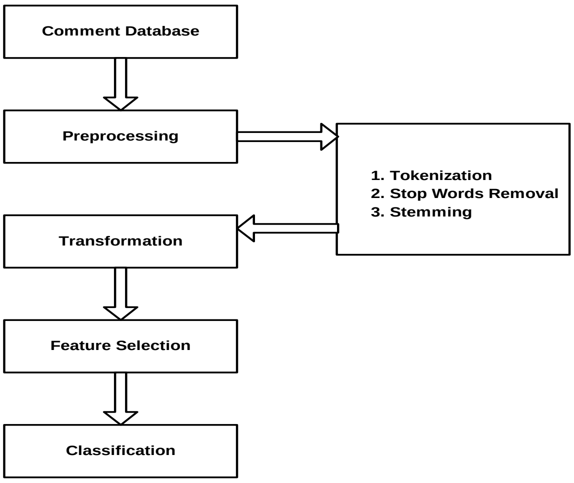


Fig.5 Steps and Techniques involved in Sentimental Analysis

Transformation of documents into TF-IDF:

The weight of each word in the corpus is calculated with the help of TF-IDF, so that it is easy to determine what words in the corpus of documents might be more favorable to use in a further processing. TF-IDF calculates [9] values for each word in a document defined as below:

$$wd = fw,d* \log(|D|fw,D)$$

D is collection of documents ,w represents words, d is individual document belongs to D,|D| is size of corpus, fw,d is number of times w appears in d,fw,D is number of documents in which w occurs in D.

Feature Vector Extraction:

Sentiment tokens and sentiment scores are information extracted from the original dataset. They are also known as features, which will be used for sentiment categorization. In order to train the classifiers, each entry of training data needs to be transformed to a vector that contains those features, namely a feature vector. For the sentence-level (review-level) categorization, a feature vector is formed based on a sentence (review).One challenge is to control each vector's dimensionality. The challenge is actually twofold: Firstly, a vector should not contain an abundant amount (thousands or hundreds) of features or values of a feature, because of the curse of dimensionality; secondly, every vector should have the same number of dimensions, in order to fit the classifiers. This challenge, particularly applies to sentiment tokens: On one hand, there are 11,478 word tokens as well as 3,023 phrase tokens; On the other hand, vectors cannot be formed by simply including the tokens appeared in a sentence (or a review), because different sentences (or reviews) tend to have different, amount of tokens, leading to the consequence that the generated vectors are in different dimensions. Since we are only concern each sentiment token's appearance inside a sentence or a review, to overcome the challenge, two binary strings are used to represent each token's appearance. One string with 11,478 bits is used for word tokens, while the other one with a bit-length of 3,023 is applied for phrase tokens.

Classification:

The corpus contains 1.6million machine-tagged Twitter messages. Each message is tagged based on the emoticons (_as positive, _as negative) discovered inside the message. The aforementioned flaws have been somewhat overcome in the following two ways: First, each product review receives inspections before it can be posted. Second, each review must have a rating on it that can be used as the ground truth. The rating is based on a star-scaled system, where the highest rating has 5 stars and the lowest rating has only 1 star.

Methods

Data collection

Data used in this paper is a set of product reviews collected from amazon.com. From February to April 2014, we collected, in total, over 5.1 millions of product review in which the products belong to 4 major categories: beauty, book, electronic, and home. Those online reviews were posted by over 3.2 millions of reviewers (customers) towards 20,062 products. Each review includes the following information: 1) reviewer ID; 2) product ID; 3) rating; 4) time of the review; 5) helpfulness; 6) review text.

Sentiment sentences extraction and POS tagging

A sentiment sentence is the one that contains, at least, one positive or negative word. All of the sentences were firstly tokenized into separated English words. Every word of a sentence has its syntactic role that defines how the word is used. The syntactic roles are also known as the parts of speech. There are 8 parts of speech in English: the verb, the noun, the pronoun, the adjective, the adverb, the preposition, the conjunction, and the interjection. In natural language processing, Part Of Speech (POS) taggers have been developed to classify words based on their parts of speech. For sentiment analysis, a POS tagger is very useful because of the following two reasons: 1) Words like nouns and pronouns usually do not contain any sentiment. It is able to filter out such words with the help of a POS tagger; 2) A POS tagger can also be used to distinguish words that can be used in different parts of speech. For instance, as a verb, "enhanced" may conduct different amount of sentiment as being of an adjective. Each sentence was then tagged using the POS tagger. Given the enormous amount of sentences, a Python program that is able to run in parallel was written in order to improve the speed of tagging. As a result, there are over 25 million adjectives, over 22 million adverbs, and over 56 million verbs tagged out of all the sentiment sentences, because adjectives, adverbs, and verbs are words that mainly convey sentiment.

Negation phrase identification

Words such as adjectives and verbs are able to convey opposite sentiment with the help of negative prefixes. For instance, consider the following sentence that was found in an electronic device's review: "The built in speaker also has its uses, but so far nothing revolutionary." The word, "revolutionary" is a positive word. SVM- Support vector machines are universal learners. Remarkable property of SVM is that their ability to learn can be independent of dimensionality of feature space. SVM measures the complexity of Hypothesis based on margin that separates the plane and not number of features.

SVM learning Algorithms for Text Categorization

SVM has defined input and output format. Input is a vector space and output is 0 or 1 (positive/negative). Text document in original form are not suitable for learning. They are transformed into format which matches into input of machine learning algorithm input. For this preprocessing on text documents is carried out. Then we carryout transformation. Each word will correspond to one dimension and identical words to same dimension. As mentioned before we will see TF-IDF for this purpose. Now a machine learning algorithm is used for learning how to classify documents, i.e. creating a model for input-output mappings.SVM has been proved one of the powerful learning algorithm for text categorization,

SVM based Evaluation

Text categorization systems may make mistakes. To compare different text classifiers for deciding which one is better, performance measures are used. Some of these measures the performance on one binary category, others aggregate per-category measures, to give an overall performance. TP, FP, TN, FN are the number of true/false positives/ negatives. We have a data set with 10.000 text messages where the model correctly predicts 9.700 negative messages, and 100 positive messages. The model still incorrectly predicts 150 messages which are positive to be negative, and 50 messages which are negative to be positive. The resulting Confusion Matrix [Table. 1] is shown below.

Table. 1 Confusion Matrix on Sentiment classification task

	Positive (Predicted)	Negative (Predicted)
Positive (Actual)	100	50
Negative (Actual)	150	9700

For the binary classification problems, which was our case situation , we can derive from those metrics two equations called sensitivity and specificity. They are commonly used for the evaluation of any binary classifier. The Specificity (TNR) measures the proportion of messages that are negative (TN) of all the messages that are actually negative (TN+FP). It can be looked at as the probability that the message is classified as negative given that the message does not contain negative words. With higher specificity, fewer positive messages are labeled as negative. On the other hand, Sensitivity (TPR) is the proportion of messages that are positive (TP) of all the messages that are actually positive (TP+FN). It can be seen as the probability that the message is positive given that the patient contain positive words. With higher sensitivity, fewer actual messages will be classified as negative.

Sensitivity can be expressed as

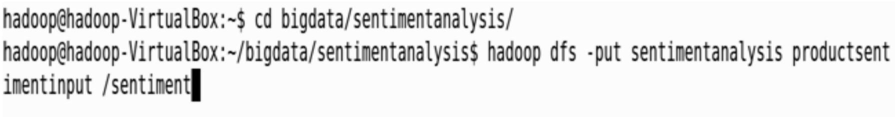
- TP / (TP+FN)

and then Specificity which is

- TN / (TN+FP) In general here, Sensitivity means the accuracy on the class Negative, and Specificity means the accuracy on the class Positive. So using these metrics, what is the accuracy on Positive and Negative messages.
- Sensitivity = TP / (TP+FN) = 100/(100+50) = 0.4 = 40%
- Specificity = TN / (TN+FP) = 9700/(9700+150) = 0.98 = 98%

RESULTS AND DISCUSSION

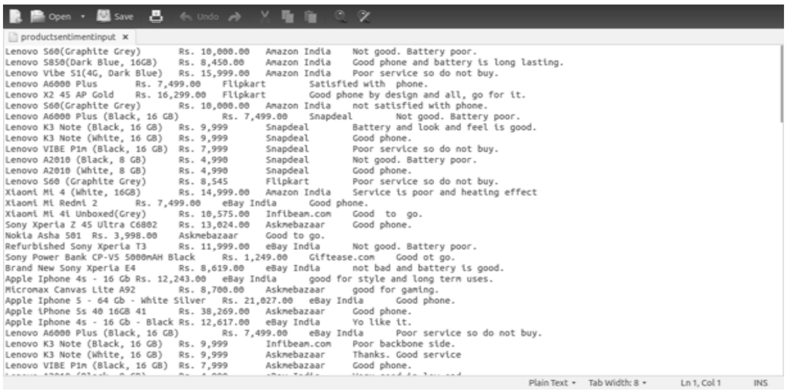
The sentimental analysis of the product review data performed at map reduce architecture using big data framework. The experiment is carried out with the Amazon customer review datasets around 62,250 of mobile product reviews. The graphs are plotted through r programming.



```
hadoop@hadoop-VirtualBox:~$ cd bigdata/sentimentanalysis/
hadoop@hadoop-VirtualBox:~/bigdata/sentimentanalysis$ hadoop dfs -put sentimentanalysis productsent
imentinput /sentiment
```

Fig.6 Sentiment Directory

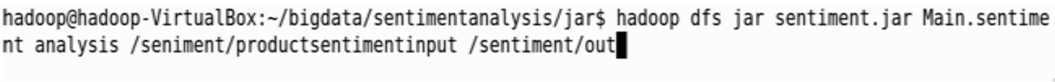
Fig. 6 shows the sentiment directory is created in the Hadoop framework and input is loaded into the directory named product sentiment input.



```
Lenovo S60(Graphite Grey)       Rs. 10,000.00   Amazon India    Not good. Battery poor.
Lenovo S850(Dark Blue, 16GB)    Rs. 8,450.00    Amazon India    Good phone and battery is long lasting.
Lenovo Vibe S1(4G, Dark Blue)   Rs. 15,999.00   Amazon India    Poor service so do not buy.
Lenovo A6000 Plus       Rs. 7,499.00    Flipkart        Satisfied with  phone.
Lenovo X2 45 AP Gold    Rs. 16,299.00   Flipkart        Good phone by design and all, go for it.
Lenovo S60(Graphite Grey)       Rs. 10,000.00   Amazon India    not satisfied with phone.
Lenovo A6000 Plus (Black, 16 GB)        Rs. 7,499.00    Snapdeal        Not good. Battery poor.
Lenovo K3 Note (Black, 16 GB)   Rs. 9,999       Snapdeal        Battery and look and feel is good.
Lenovo K3 Note (White, 16 GB)   Rs. 9,999       Snapdeal        Good phone.
Lenovo VIBE P1m (Black, 16 GB)  Rs. 7,999       Snapdeal        Poor service so do not buy.
Lenovo A2010 (Black, 8 GB)      Rs. 4,990       Snapdeal        Not good. Battery poor.
Lenovo A2010 (White, 8 GB)      Rs. 4,990       Snapdeal        Good phone.
Lenovo S60 (Graphite Grey)      Rs. 8,545       Flipkart        Poor service so do not buy.
Xiaomi Mi 4 (White, 16GB)       Rs. 14,999.00   Amazon India    Service is poor and heating effect
Xiaomi Mi Redmi 2       Rs. 7,499.00    eBay India      Good phone.
Xiaomi Mi 4i Unboxed(Grey)      Rs. 10,575.00   Infibeam.com    Good  to  go.
Sony Xperia Z 45 Ultra C6802    Rs. 13,024.00   Askmebazaar     Good phone.
Nokia Asha 501  Rs. 3,998.00    Askmebazaar     Good to go.
Refurbished Sony Xperia T3      Rs. 11,999.00   eBay India      Not good. Battery poor.
Sony Power Bank CP-V5 5000mAH Black     Rs. 1,249.00    Giftease.com    Good ot go.
Brand New Sony Xperia E4        Rs. 8,619.00    eBay India      not bad and battery is good.
Apple Iphone 4s - 16 Gb Rs. 12,243.00   eBay India      good for style and long term uses.
Micromax Canvas Lite A92        Rs. 8,700.00    Askmebazaar     good for gaming.
Apple Iphone 5 - 64 Gb - White Silver   Rs. 21,027.00   eBay India      Good phone.
Apple iPhone 5s 40 16GB 41      Rs. 38,269.00   Askmebazaar     Good phone.
Apple Iphone 4s - 16 Gb - Black Rs. 12,617.00   eBay India      Yo like it.
Lenovo A6000 Plus (Black, 16 GB)        Rs. 7,499.00    eBay India      Poor service so do not buy.
Lenovo K3 Note (Black, 16 GB)   Rs. 9,999       Infibeam.com    Poor backbone side.
Lenovo K3 Note (White, 16 GB)   Rs. 9,999       Askmebazaar     Thanks. Good service
Lenovo VIBE P1m (Black, 16 GB)  Rs. 7,999       Askmebazaar     Good phone.
```

Fig. 7 Product Sentiments

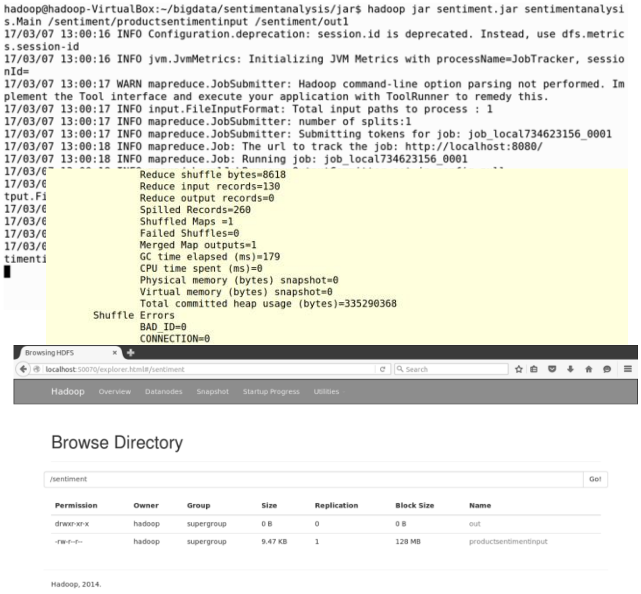
Fig. 7 shows the product sentiments from the customer collected from Amazon database and the input consist of product details and the corresponding sentiments.



```
hadoop@hadoop-VirtualBox:~/bigdata/sentimentanalysis/jar$ hadoop dfs jar sentiment.jar Main.sentime
nt analysis /seniment/productsentimentinput /sentiment/out
```

Fig. 8 Job Submission in Hadoop

Fig. 8 shows the job submission in hadoop framework with the corresponding input and output destinations.



```
hadoop@hadoop-VirtualBox:~/bigdata/sentimentanalysis/jar$ hadoop jar sentiment.jar sentimentanalysi
s.Main /sentiment/productsentimentinput /sentiment/out1
17/03/07 13:00:16 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metric
s.session-id
17/03/07 13:00:16 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessio
nId=
17/03/07 13:00:17 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Im
plement the Tool interface and execute your application with ToolRunner to remedy this.
17/03/07 13:00:17 INFO input.FileInputFormat: Total input paths to process : 1
17/03/07 13:00:17 INFO mapreduce.JobSubmitter: number of splits:1
17/03/07 13:00:17 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local734623156_0001
17/03/07 13:00:18 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
17/03/07 13:00:18 INFO mapreduce.Job: Running job: job_local734623156_0001
17/03/0
17/03/0
tput.Fi
17/03/0
17/03/0
17/03/0
17/03/0
timenti
```

```
                Reduce shuffle bytes=8618
                Reduce input records=130
                Reduce output records=0
                Spilled Records=260
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=179
                CPU time spent (ms)=0
                Physical memory (bytes) snapshot=0
                Virtual memory (bytes) snapshot=0
                Total committed heap usage (bytes)=335290368
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
```

Browse Directory

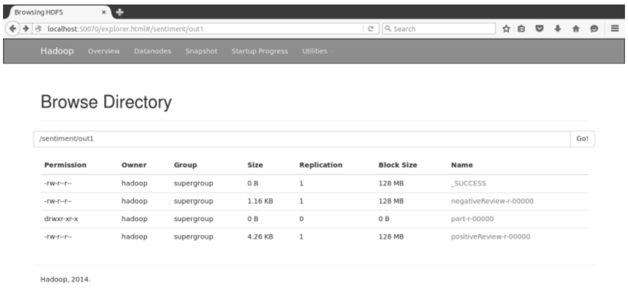
/sentiment

Permission	Owner	Group	Size	Replication	Block Size	Name
drwxr-xr-x	hadoop	supergroup	0 B	0	0 B	out
-rw-r--r--	hadoop	supergroup	9.47 KB	1	128 MB	productsentimentinput

Hadoop, 2014.

Fig. 9 Map Reduce Process

Fig. 9 shows successful completions of map reduce process and the output is written back into the corresponding destination.



Browse Directory

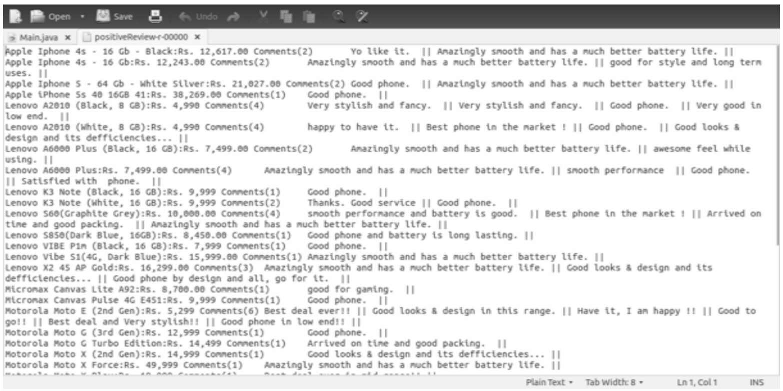
/sentiment/out1 Go!

Permission	Owner	Group	Size	Replication	Block Size	Name
-rw-r--r--	hadoop	supergroup	0 B	1	128 MB	_SUCCESS
-rw-r--r--	hadoop	supergroup	1.16 KB	1	128 MB	negativeReview-r-00000
drwxr-xr-x	hadoop	supergroup	0 B	0	0 B	part-r-00000
-rw-r--r--	hadoop	supergroup	4.26 KB	1	128 MB	positiveReview-r-00000

Hadoop, 2014.

Fig. 10 Positive and Negative Sentiments Classification

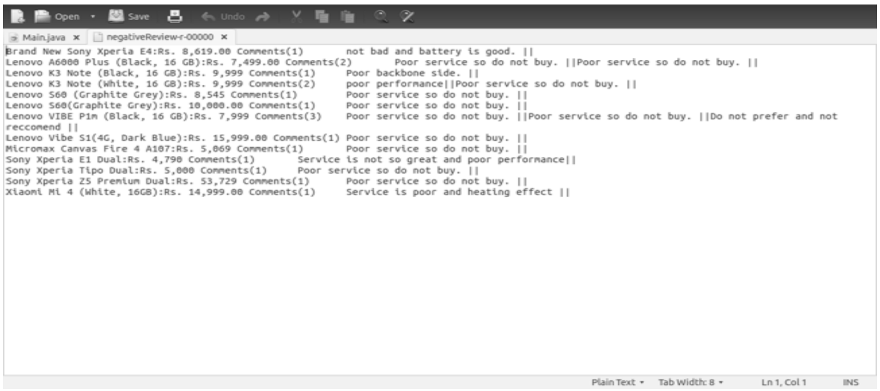
Fig. 10 shows the successful completion of classification of user product reviews based on splitting negative and positive sentiments.



```
Apple Iphone 4s - 16 Gb - Black:Rs. 12,617.00 Comments(2)       Yo like it.  || Amazingly smooth and has a much better battery life. ||
Apple Iphone 4s - 16 Gb:Rs. 12,243.00 Comments(2)       Amazingly smooth and has a much better battery life. || good for style and long term
uses. ||
Apple Iphone 5 - 64 Gb - White Silver:Rs. 21,027.00 Comments(2) Good phone.  || Amazingly smooth and has a much better battery life. ||
Apple iPhone 5s 40 16GB 41:Rs. 38,269.00 Comments(1)    Good phone.  ||
Lenovo A2010 (Black, 8 GB):Rs. 4,990 Comments(4)        Very stylish and fancy.  || Very stylish and fancy.  || Good phone.  || Very good in
low end.  ||
Lenovo A2010 (White, 8 GB):Rs. 4,990 Comments(4)        happy to have it.  || Best phone in the market ! || Good phone.  || Good looks &
design and its defficiencies... ||
Lenovo A6000 Plus (Black, 16 GB):Rs. 7,499.00 Comments(2)       Amazingly smooth and has a much better battery life. || awesome feel while
using. ||
Lenovo A6000 Plus:Rs. 7,499.00 Comments(4)      Amazingly smooth and has a much better battery life. || smooth performance  || Good phone.
|| Satisfied with  phone.  ||
Lenovo K3 Note (Black, 16 GB):Rs. 9,999 Comments(1)     Good phone.  ||
Lenovo K3 Note (White, 16 GB):Rs. 9,999 Comments(2)     Thanks. Good service || Good phone.  ||
Lenovo S60(Graphite Grey):Rs. 10,000.00 Comments(4)     smooth performance and battery is good.  || Best phone in the market ! || Arrived on
time and good packing.  || Amazingly smooth and has a much better battery life. ||
Lenovo S850(Dark Blue, 16GB):Rs. 8,450.00 Comments(1)   Good phone and battery is long lasting. ||
Lenovo VIBE P1m (Black, 16 GB):Rs. 7,999 Comments(1)    Good phone.  ||
Lenovo Vibe S1(4G, Dark Blue):Rs. 15,999.00 Comments(1) Amazingly smooth and has a much better battery life. ||
Lenovo X2 45 AP Gold:Rs. 16,299.00 Comments(3)  Amazingly smooth and has a much better battery life. || Good looks & design and its
defficiencies... || Good phone by design and all, go for it.  ||
Micromax Canvas Lite A92:Rs. 8,700.00 Comments(1)       good for gaming.  ||
Micromax Canvas Pulse 4G E451:Rs. 9,999 Comments(1)     Good phone.  ||
Motorola Moto E (2nd Gen):Rs. 5,299 Comments(6) Best deal ever!! || Good looks & design in this range. || Have it, I am happy !! || Good to
go!! || Best deal and Very stylish!! || Good phone in low end!! ||
Motorola Moto G (3rd Gen):Rs. 12,999 Comments(1)        Good phone.  ||
Motorola Moto G Turbo Edition:Rs. 14,499 Comments(1)    Arrived on time and good packing.  ||
Motorola Moto X (2nd Gen):Rs. 14,999 Comments(1)        Good looks & design and its defficiencies... ||
Motorola Moto X Force:Rs. 49,999 Comments(1)    Amazingly smooth and has a much better battery life. ||
```

Fig. 11 Extraction of Positive Sentiments

Fig. 11 shows the positive sentiments of mobile devices with their brand name and ratings extracted using POS method.



```
Brand New Sony Xperia E4:Rs. 8,619.00 Comments(1)       not bad and battery is good. ||
Lenovo A6000 Plus (Black, 16 GB):Rs. 7,499.00 Comments(2)       Poor service so do not buy. ||Poor service so do not buy. ||
Lenovo K3 Note (Black, 16 GB):Rs. 9,999 Comments(1)     Poor backbone side. ||
Lenovo K3 Note (White, 16 GB):Rs. 9,999 Comments(2)     poor performance||Poor service so do not buy. ||
Lenovo S60 (Graphite Grey):Rs. 8,545 Comments(1)        Poor service so do not buy. ||
Lenovo S60(Graphite Grey):Rs. 10,000.00 Comments(1)     Poor service so do not buy. ||
Lenovo VIBE P1m (Black, 16 GB):Rs. 7,999 Comments(3)    Poor service so do not buy. ||Poor service so do not buy. ||Do not prefer and not
reccomend ||
Lenovo Vibe S1(4G, Dark Blue):Rs. 15,999.00 Comments(1) Poor service so do not buy. ||
Micromax Canvas Fire 4 A107:Rs. 5,069 Comments(1)       Poor service so do not buy. ||
Sony Xperia E1 Dual:Rs. 4,790 Comments(1)       Service is not so great and poor performance||
Sony Xperia Tipo Dual:Rs. 5,000 Comments(1)     Poor service so do not buy. ||
Sony Xperia Z5 Premium Dual:Rs. 53,729 Comments(1)      Poor service so do not buy. ||
Xiaomi Mi 4 (White, 16GB):Rs. 14,999.00 Comments(1)     Service is poor and heating effect ||
```

Fig. 12 Extraction of Negative Sentiments

Fig. 12 shows the negative sentiments of mobile devices with their brand name and ratings extracted using POS method.

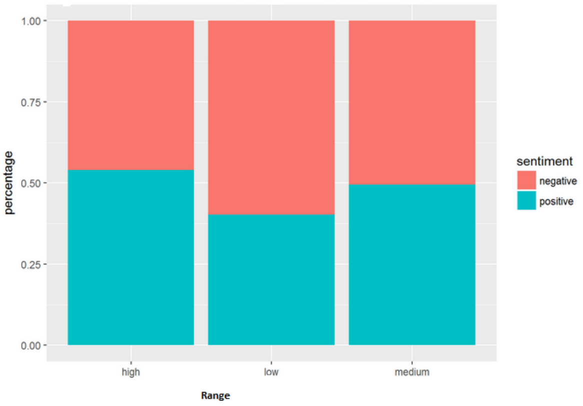


Fig. 13 Polarity Plot based on Sentiments

Fig. 13 shows the polarity plot for mobiles based on sentimental score of customer's reviews for positive and negative sentiments.

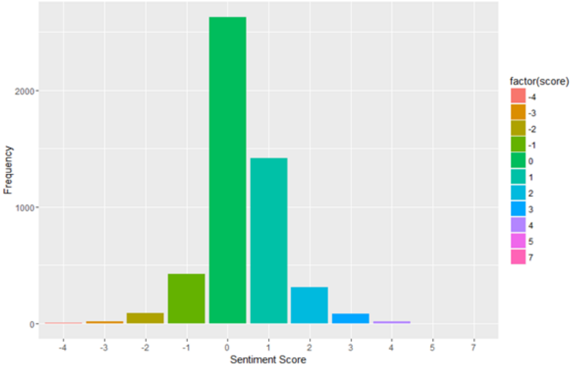


Fig. 14 Customer Sentimental Score

Fig.14 shows the sentimental score based on ratings and textual information from the customers for mobile phones. The below Fig. 15 shows the ROC curve for SVM based prediction model for classification of positive and negative labeled sentiments.

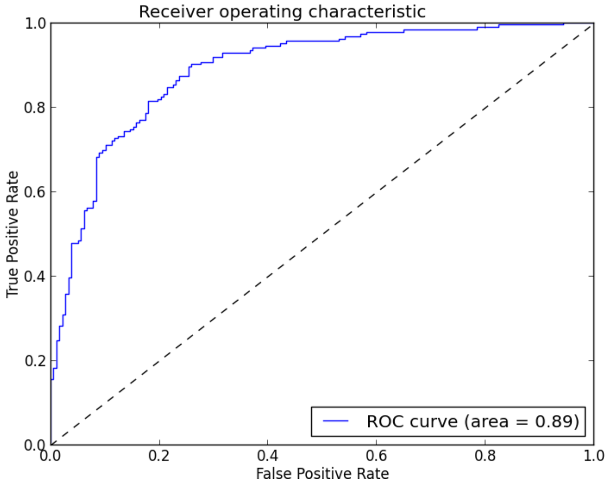


Fig. 15 Roc Curve for SVM based Prediction

CONCLUSION AND FUTURE WORK

Sentiment analysis or opinion mining is a field of study that analyzes people's sentiments, attitudes, or emotions towards certain entities. This paper handles a fundamental problem of sentiment analysis, sentiment polarity categorization. Online product reviews from Amazon.com are selected as data used for this study. A sentiment polarity categorization process has been proposed along with detailed descriptions of each step. Experiments for both sentence-level categorization and review-level categorization have been performed.

REFERENCES

1. *Kim, Soo-Min, and Eduard Hovy, "Determining the sentiment of opinions." Proceedings of the 20th international conference on Computational Linguistics, Association for Computational Linguistics, pp. 1-8, 2004.*
2. *Liu, Bing, Minqing Hu, and Junsheng Cheng. "Opinion observer: analyzing and comparing opinions on the web", Proceedings of the 14th international conference on World Wide Web, pp. 342-351, ACM, 2005.*
3. *Pak, Alexander, and Patrick Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", In LREc. Vol. 10. No. 2010, pp. 1320-1326, 2010.*
4. *Choi, Yejin, and Claire Cardie, "Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification", Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp.590-598, 2009.*
5. *Sarvabhotla, Kiran, Prasad Pingali, and Vasudeva Varma, "Sentiment classification: a lexical similarity based approach for extracting subjectivity in documents", Information Retrieval, Vol. 14(3), pp. 337-353, 2011.*
6. *Chesley, P., Vincent, B., Xu, L., & Srihari, R. K, "Using verbs and adjectives to automatically classify blog sentiment", Training, 580(263), 233, 2010.*
7. *Bhatt, A., Patel, A., Chheda, H., & Gawande, K, "Amazon Review Classification and Sentiment Analysis", International Journal of Computer Science and Information Technologies, Vol. 6(6), pp.5107-5110, 2015.*
8. *Rain, Callen, "Sentiment Analysis in Amazon Reviews Using Probabilistic Machine Learning", Swarthmore College, (2013).*
9. *Mochamad Wahyudi and Dinar Ajeng Kristiyanti, "Sentiment Analysis of Smartphone Product Review Using Support Vector Machine Algorithm-Based Particle Swarm Optimization", Journal of Theoretical & Applied Information Technology, Vol. 91(1), pp. 189-201, 2016.*
10. *Jalpa Mehta, Jayesh Patil, Rutesh Patil, Mansi Somani, Sheel Varma, "Sentiment Analysis on Product Reviews using Hadoop", International Journal of Computer Applications, Vol. 142(11), pp. 38-41, 2016.*
11. *Madhura G K, Prof.Shivamurthy R C, " Twitter Sentiment Analysis for Product Reviews to Gather Information using Machine Learning Technique", International Journal of Research In Science & Engineering, Vol. 1(2), pp. 1-6.*
12. *Asghar, M. Z., Khan, A., Ahmad, S., & Kundi, F. M, "A review of feature extraction in sentiment analysis", Journal of Basic and Applied Scientific Research, Vol. 4(3), pp.181-186, 2014.*
13. *Feczko, M., Schaye, A., Marcus, M., & Nenkova, A, "Sentisummary: Sentiment summarization for user product reviews", In proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Vol. 1, pp. 265-271, 2008.*
14. *Schouten, K., & Frasincar, F, "Finding implicit features in consumer reviews for sentiment analysis", In International Conference on Web Engineering, Springer International Publishing, pp. 130-144, 2014.*
15. *Hadoop(The Definitive Guide) By Tom White, 2012.*
16. *Book Analytics with R and Hadoop By Vignesh Prajapati, 2013*