SENTIMENT ANALYSIS OF SOCIAL MEDIA DATA

[1] DENSY DAVID E, [2] ANVER S R

[1] Mtech Scholar, [2] Associate Professor
[1,2] *Department of Computer Science and Engineering*
[1, 2] *LBS Institute of Technology for Woman Trivandrum*, Kerala, India.

Abstract : The amount of data in our industry and the world is exploding. Data is being collected and stored athuge rates. The challenge is not only to store and manage thevast volume of data ("Big Data"), but also to analyze and extract meaningful value from it. Big Data comes in many forms. It comes as customer information and transactions contained in customer-relationship management and enterprise resource planning systems, HTML based web stores etc. The information posted on social network and blogging sites including Facebook, Twitter and Youtube mainly constitutes the social media data, which is a form of Big Data. When purchasing a product for the first time one usually needs to choose among several products with similar characteristics. As the companies promote their brands and products by pointing out only good characteristics, the best way to choose the most suitable product is to rely upon the opinions of others.ie., the independent unbiased consumer reviews are known to be the most credible sources of product or service information and people tend to rely primarily on them when making a decision about a purchase. The system collects opinions about an item from the web, evaluate them aggregates these evaluations and offers cumulative, east-to-understand information. Generated information is intended for the possible prospective customer but also for the company managers providing them with additional guidancein future business development.

IndexTerms - Big Data, Sentiment Analysis, Hadoop,HDFS, MapReduce, Twitter.

I. INTRODUCTION

Our day-to-day life depends much on what other people think and say, ie; on their opinions. In other words opinions play a major role in our life while we are in a position to take a decision on any matter. Opinions on anysubject/object can be positive or negative. It is our duty tofind out the benefits from these opinions. While we are purchasing a product for the first time we do not know much about it. In such a situation, we enquire about the product to our friends and we select the product accordingto the opinions we got from others. There comes the importance of opinion mining/retrieval in our life. Weneed to select the best product, so the opinion mining process has to be efficient in all means.

Everybody can express their opinion on any subject/object. Earlier days it was through mouth-to mouth or in articles. But with the rapid development of web, anyone can express their thinkings and opinion through web, ie; we can say that the web has emotions[3].Users can post their reviews or opinions in the web through Blogs, User comments, Review websites, Community websites, etc. There are two main types of textual information - facts and opinions. A fact is a universally accepted data whereas an opinion is something that is mainly based on a user‟s perspective towards that subject/object. Current search engines searchfor facts not for opinions and the current search ranking strategy is not appropriate for opinion retrieval/search.There are mainly two types of opinions – direct opinions and comparisons. The sentiment expressions on some subject or object can be regarded as direct opinions whereas the comparison is the relations expressingsimilarities or differences of more than one object. The basic components of an opinion are – opinion holder,object and the opinion.

A large amount of data is being produced in theweb daily. We need to efficiently manage these data foruse. As the amount and size of data increases, it isdifficult to manage and coordinate them. The largeamount of data in the web is considered as the "BigData". It can be defined as the volume of data thattraditional database methods and tools could not process efficiently. As the amount of data is huge, the actionabledata from various data sources has to be extracted. TheBig Data can be considered based on four basicdimensions – Volume, Velocity, Variety and Veracity [1].Volume is a measure of amount of data, Velocity is a measure of speed of data being produced, Veracity is themeasure of difference in types of data produced (ie; data can be text, image, audio, video etc) and Veracity is ameasure of uncertainty of data.

Here we are considering about the product review mining, ie; What features of the product do customers like and which do they dislike?. We are performing review mining in data from Twitter. Twitter is a popular microblogging website where users have theability to send/post updates. The posts in Twitter is calleda 'tweet'. A tweet is 140 characters in length which canbe send to a group of friends(followers) and express atweeter's emotion on a particular subject/object. Tweetsby default are public, which permits people to followothers and read each other's tweets without giving mutualpermission. However, senders can restrict delivery tofriends only. Our aim is to analyze the hidden sentimentin a tweet. A sentiment can be defined as a thought, view,or attitude, especially one based mainly on emotioninstead of reason. Sentiment Analysis/opinion mining isdefined as a Natural Language Processing andInformation Extraction task which aims to determine theattitude of a speaker or a writer. It identifies the phrases ina text that bears some sentiment and the subject towardswhom the sentiment is directed. Sentiment Analysis canbe classified into three – Document based, Sentence basedand Phrase based – dealing with extracting the sentimentfrom a document as a whole, a sentence and a phraserespectively. The sentiment analysis leads to a three-wayclassification of the sentiments – positive, negative andneutral. In the process each term/phrase is given asentiment score and the average score of a document/sentence/phrase is being calculated.Conclusions regarding the hidden sentiment about aproduct is being derived based on these calculated scores. The collected reviews/tweets and the results produced arestored and managed efficiently using HDFS (HadoopDistributed File System). HDFS is designed for reliablestorage and management of very large files that areunstructured. It supports write-once-read-many semantics on very large unstructured files. HDFS uses the column-oriented database HBase and MapReduce technique fordata storage. Thus, the system analyzes the hiddensentiment in tweets regarding product review andprovides results in a structured format.

The major problem faced by customers during purchasing is the selection of suitable and best product. Our aim is to develop a method for automatic sentiment analysis of Twitter messages for providing reviews orservices to consumers who are trying to know or inquire about a product or service. In the current system. Sentimental analysis is done based on the reviews of customers in websites. Here the comments may or may not be updated recently. Suppose a product is good at the start and not as much good recently the analysis will be based on the old comments. So current updates may notbe revealed. So we need a new system as a solution. Sincesocial networking is becoming a boon in the current instance and twitter and facebook showing an important role, Reviews from these pages are all of huge importance. So we here by provide a new idea of data mining based on twitter comments and reviews. So sentimental analysis will become more reliable and efficient.

II. LITERATURE SURVEY

A. LANGUAGE SPECIFIC AND TOPIC FOCUSSED WEB CRAWLING

The system performs efficient crawling technique with a common text classification tool[2]. For most corpus basedapproaches to NLP document collections in a specificlanguage that cover a particular domain are required. These collections are used as training data to build accurate statistical models reflecting their characteristics, constrained by the language, theme and style. Local Web collections are usually created by crawling the WWW starting with few seed URLs.

The crawler is a program that fetches each webpage, follows its outgoing links and repeats fetching process recursively. Focused crawling implies fetchingonly those pages that are relevant to a particular topic or language. Different approaches were proposed for the focusing strategy. In general, they all are based on the assumption that webpages on a given topic are more likely to link to those on the same topic. Starting from a set of webpages that represent the given topic, the crawlerfollows their links and is restricted by either content words on the outgoing webpages, or the graph structure ofthe Web, or URL tokens, or the combination of these criteria. They are used to prevent crawling of unrelated websites which would result in a deviation from the specific topic. Two parameters to be considered for efficient focused crawling are selection of good starting points for crawling and content analysis of fetched webpages.

A context focused crawler outperforms thestandard focused crawler in terms of required steps till atopic specific webpage is found. However, the approachdepends on the nature of a category. It is only useful for categories that have a standard way of hierarchicalpositioning on the web. The current attempts in focusedcrawling do not pay enough attention to the actual content of their training and test collections. They briefly mentionconsidering content words appearing on the webpages astopic filters. Some use TFxIDF weightening schemes torestrict these sets to most representative words and speedup the classification process. Some approaches payextensive attention to collecting focused seed queries intheir approach. They use queries to acquire languagespecific corpora for minority languages such as Slovenianand Tagalog. The queries are automatically constructedfrom terms with the highest probability scores, computedfrom two sets with relevant and nonrelevant documents. New documents retrieved with these queries are thenadditionally categorized with the text classification tool toincrease these initial sets. The disadvantage of thisapproach is the need of large contrast corpus and, mostimportant, the absence of the crawling element.

The content based focused crawling is performedin two steps. Firstly, a list with topic and language specific seed URLs are created. Secondly, a open-source crawler is runned starting from these URLs and use the text categorization tool to avoid crawling of irrelevantwebpages. In order to collect seed URLs for the initialization of the crawling process, first domain specificqueries are needed. Given a document collection, a two tofive words window to extract all possible phrasesconsisting of non stopwords that appear in these documents are used. For each phrase *i* in samplecollection C, compute its average TF X $IDF_{i,c}$ value. where $TF_{i,c}$ is term frequency, i.e. the frequency of aphrase i in the document c; DF_i is document frequency, i.e. the number of documents in the collection that containthe phrase i; and |C| is the number of documents in the collection. The top phrases ranked according their TF X IDF value, will be already domain specific. The resulting phrases are used as focused queries and acquire seed URLs by sending these queries to a standard web search engine. To create domain specific models two different document collections for each language are used. First, using all documents of the sample collection. Second,with a document collection from a different domain, which consisted of news articles downloaded from the Internet. Each time a webpage is fetched by the crawler, Text categorization tool is used twice to ensure its similarity to the sample collection. The webpage is only preserved if it passes both tests.

B. AUTOMATICALLY EXTRACTING USER REVIEWS FROM FORUM SITES.

User reviews in forum sites are the important information source for many popular applications (e.g., monitoring and analysis of public opinion), which are usually represented in form of structured records[3]. With therapid development of Web, web users can freely post theirreviews for specific events or objects on webpages tovoice their opinions. The number of reviews is increasing at a surprising speed, and the reviews cover almost all the domains in the real world, such as commerce, politics, and entertainment. Besides the variety of web page templates, user generated reviews raise two new challenges. First, the inconsistency of review contents in terms of both the document object model (DOM) tree and visual appearance impair the similarity between review records; second, the review content in a review record corresponds to complicated subtrees rather than singlenodes in the DOM tree. The system consists of two subtasks: review record extraction and review content extraction. Review record extraction detects the boundaries of the review records embedded in web pages and further extracts them as the input of the next subtask. Review content extraction extracts the review contents from the extracted review records. The major componentsinclude following tasks.

Web page representation: A review page is parsed into a DOM tree, and the visual information of the tree nodes is attached.

Review record extraction: The minimum subtree that contains all review records is detected from the DOMtree first, and then all review records are extracted through removing the noise and detecting the boundaries of the review records.

Review content extraction: For review content extraction, the minimum subtree that contains the pure review content of each review record is extracted.

In webpage representation we parse a review page into a DOM tree, and perform the extraction task in the DOM tree. Due to the poor ability of HTML tags withrespect to semantic representation, we combine the visual information into the DOM tree to improve the extraction performance. Visual information on web pages has proven to be a very useful feature for web data search andextraction. During the parsing process, useful visual information is obtained and attached to the nodes of the DOM tree. The visual information used in this paper is classified into three types – Position (the coordinate of theleft-top corner of a node), size (the width and height of the rectangle that a node occupies in the web page) and Font (the fonts of the texts of a node, including font size, font style, and font color). Based on observations of a large number of review pages, a group of useful featuresis generalized for web review extraction. The process of review record extraction includes three steps: first,Tregion is detected in the DOM tree; next, the noise is removed from T_{region}, finally, the boundaries of the reviewrecords are detected, i.e., how many sequential subtrees makeup one review record. The process of detecting Tregion is very similar to that of detecting search result records group in search result pages. The problem is solved by using four simple visual features of the search result records group: it occupies a large area; it is centrally located; it contains many characters; and it has a large number of records. This method is used for detecting T_{region}. Because review records are ordered according to their post-dates, we supplement a new feature to improve the accuracy: it contains many ordered dates.

Review content extraction deals with how toextract the pure review content from a review record. Dueto the diversity of review contents, they are very inconsistent in terms of both tree structure and text length.The method for review content extraction is based on the inconsistencies of tree structure and text length. First, we identify the T_{review} containing review content and then extracting the minimum subtree containing review content. The review content in a review record is one complicated subtree rather than one single node in the DOM tree. Therefore, the review content extraction process is actually to extract the minimum subtrees (denoted as Tcontent) that contain the pure review contents. This is known as Direct extraction, ie., the review pages from different forum sites are inputted indiscriminately, but every page must contain multiple review records.

In Wrapper-based extraction, the review record wrapper and the review content wrapper are generated by employing direct extraction with one sample page. Then, the review records and the review contents are extracted in turn with their corresponding wrappers.

C. SENTIMENT CLASSIFICATION USING MACHINE LEARNING TECHNIQUES

Various supervised or data-driven techniques to Sentiment Analysis includes Naïve Byes, Maximum Entropy, SVM etc[4].

Support Vector Machines: Support vector machines (SVMs) belong to the family of linear classifiers and creates feature-vector-based classifiers. The purpose of linear classification is to search for alinear hyperplane in a feature space dividing all entities into two classes. The basic idea of SVMs is to search for aseparating hyperplane that has the maximum distance from the points nearest to it in the feature space. In the case of linearly separable sample, search for a hyperplane can be written down as an optimization problem. Trainingdata is used to generate a high dimensional space that can be divided by a hyperplane between positive and negativeinstances. Then new instances are classified by finding their position in the space with respect to the hyperplane.

Naive Bayes Classifier: The naive Bayesclassifier is a probabilistic classifier based on applying theBayes" theorem and (naive) statistical independence assumptions of random variables. The basic assumption isthat "the presence or absence of a particular feature is unrelated to the presence or absence of any other feature, given the class variable". The main advantage of thisclassifier is its low computational complexity and optimality, provided there is real independence of features.

Maximum Entropy Classifier: The principle of maximum entropy states that, subject to precisely stated prior data, the probability distribution which best represents the current state of knowledge is the one with largest entropy. Take a precisely stated prior data or testable information about a probability distributionfunction. Consider the set of all trial probability distributions that would encode the prior data. Of those, one with maximal information entropy is the proper distribution, according to this principle.

When the feature space is small, Naïve Bayes performs better than SVM. But SVM's perform better when feature space is increased. When feature space is increased, Maximum Entropy may perform better than Naïve Bayes but it may also suffer from overfitting.

D. TWITTER AS A CORPUS FOR SENTIMENT ANALYSIS AND OPINION MINING

Twitter contains an enormous number of text posts and itgrows every day. The collected corpus can be arbitrarilylarge. The system collects a corpus and perform statisticalanalysis of collected corpus and build a sentimentclassification system for micro blogging [5]. The corpuscollection is done using Twitter API we collected acorpus of text posts and formed a dataset of three classes: positive sentiments, negative sentiments, and a set ofobjective texts (no sentiments). The two types of collectedcorpora (happy and sad) will be used to train a classifierto recognize positive and negative sentiments. In corpusanalysis, first the distribution of words frequencies in the corpus is checked and a plot of word frequencies is made.Now tag all the posts in the corpus. The collected datasetis used to extract features that will be used to train oursentiment classifier. The presence of an n-gram is used asa binary feature, while for general information retrievalpurposes, the frequency of a keyword’s occurrence is amore suitable feature, since the overall sentiment may notnecessarily be indicated through the repeated use ofkeywords. The process of obtaining n-grams from aTwitter post consists of filtering, tokenization, stop-word removal and n-gram construction. In filtering process,URL links (Eg. http://example.com), Twitter user names(eg. @alex – with symbol @ indicating a user name),Twitter special words (such as “RT”6), and emoticons areremoved. In tokenization, text is segmented by splitting itbyspaces and punctuation marks, and form a bag of words.Removing stop-words means to remove articles (‘a’,‘an’, ‘the’) from the bag of words. For constructing n-grams, a set of n-grams are made out of consecutivewords. A negation (such as ‘no’ and ‘not’) is attached toa word which precedes it or follows it. Now a sentimentclassifier is built using the multinomial Naïve Bayesclassifier. Two Bayes classifiers are trained, which usedifferent features: presence of n-grams and part-of-speech distribution information. N-gram based classifier uses thepresence of an n-gram in the post as a binary feature. Theclassifier based on POS distribution estimates probabilityof POS-tags presence within different sets of texts anduses it to calculate posterior probability. The sentimentclassifier must be able to determine positive, negative andneutral sentiments for a document.

E. USING BIG DATA AND SENTIMENT ANALYSIS IN PRODUCT EVALUATION

The system collects opinions about hotels from the web,evaluates them, aggregates these evaluations and offerscumulative, easy-to-understand information [1]. Generatedinformation is intended for the possible prospectivecustomer, but also for the hotel managers providing them with additional guidance in future business development.The goal of text mining is to derive high-qualityinformation from the text. This is typically done through recognizing the patterns in data. In other words, the purpose of text mining is to process unstructured information and to extract meaningful numeric indicesfrom it. Text analysis involves information retrieval, Natural language processing, named entity recognition, recognition of pattern identified entities, coreference, relationship, fact, and event extraction, sentiment analysis, Quantitative text analysis etc.

In Data Collection, an ideal crawler for the purpose of quickly downloading only the pages containing reviews and checking if they are updated, would be distributed or at least parallel, incremental and focused. To update a set of downloaded pages it is preferable to apply incremental crawling rather than to restart crawling. In focused crawling the space of crawled pages is narrowed by the use of a classifier, which decideswhether a page is interesting or not. The general-purpose web crawler offered by Apache Nutch, an open-source web-search software project was used. During Data Description, for each extracted review a key indicator wascreated before the review was stored into a database. The review consists of text of the review,date of the review, language of the text. Text analytics and sentiment analysiswere performed by means of the open-source software KNIME.
KNIME is a user-friendly graphical workbenchfor the entire analysis process: data access, datatransformation, initial investigation, powerful predictive analytics, visualization and reporting. By means ofKNIME a sentiment analysis stream, consisting of the following major steps, was created: retrieving data fromthe database, dictionary development and implementation,review scoring. Data were retrieved from the database inpackages of many records. Only records that were notpreviously evaluated were retrieved, assuring that thereviews from the database were evaluated only once,speeding up the whole process. A new dictionarycontaining words and phrases used in evaluation of hotel reviews was developed and implemented. It wasdeveloped based on the words and phrases found in asample of downloaded Internet hotel reviews. It wasnecessary to include the terms and phrases from thecurrent hotel reviews which could contribute to theevaluation, and to multiply them and modify them in afollowing grammar rule but including also slangexpressions and abbreviations used in everyday speech.

Review scoring is performed in following way:First every word from document is tagged according tospecification of terms and phrases in the dictionary. Thenthe document is transformed into 'Bag of words', andwords not specified within applied dictionary areexcluded. Finally, every term or phrase within dictionary has assigned category and mark, so recognized terms and phrases within review are grouped according to categoriesspecified within dictionary, counted and average mark forevery category is calculated. The results, i.e. average grade for each evaluated review for each category was written back to the Hadoop database.

F. SENTIMENT ANALYSIS ON TWEETS FOR SOCIAL EVENTS

As more and more users express their political and religious views on Twitter, tweets become valuable sources of people’s opinions. Tweets data can be efficiently used to infer people’s opinions for marketingor social studies. This paper proposes a Tweets Sentiment Analysis Model (TSAM) that can spot the societal interestand general people’s opinions in regard to a social event [6]. It can identify the positive, negative or neutral opinions and measure intensity (or strength) ofpositive/negative opinions in regard to an entity (people, organization, location, product, etc.). The conceptual framework of the TSAM consists of three modules - Feature selection module that extracts the opinionated words from each sentence, Sentiment identification module that associates expressed opinions with each relevant entity in each sentence level, and Sentiment aggregation and scoring module that calculates thesentiment scores for each entity. In feature selection of this sentiment analysis model, instead of using all thewords appearing in the news articles or tweets, we only extract the opinion-bearing words as the features to input into opinion mining algorithm. Opinion words that are primarily used to express subjective opinions in the opinion sentence are identified and extracted. Words that encode a desirable state (e.g., beautiful, awesome) have a positive orientation, while words that represent undesirable states have a negative orientation (e.g., disappointing).

The semantic orientation of a word will be used to predict the semantic orientation of each opinion sentence. Opinion sentence is a sentence that contains oneor more entities and one or more opinion words. To identify the opinionated words, we use Wilson opinion lexicon list to decide the words‟ semantic orientations. Wilson lexicon consists of three lists of subjectivity clues:

(i) the prior polarity lexicon,
(ii) the intensifier lexicon, and
(iii) the valence shifter lexicon.

In this project, only the prior polarity lexicon subjectivity clue is used. We quantify the semantic orientation of words by given each type of word a numeric score. Therefore, a positive and strong subjectivity words is assigned the semantic orientation score of +1, a positive and weak subjectivity word is assigned the semantic orientation score of +0.5, and a negative and strong subjectivity word is assignedthe semantic orientation score of −1, a negative and weak subjectivity word is assigned the semantic orientation score of −0.5, and a neutral word is given the semantic orientation score of 0. These text strings can be placed into categories (positive, negative, neutral) and one can differentiate their strength or impact by assigning different weights.

In Sentiment Analysis Technique, at first, a Sentence Sentiment Scoring Function (SSSF) is used to determine the orientation of sentiment expressed on each entity *ei* in *s* (i.e., the pair of (*ei* , *s*)). Then an Entity Sentiment Aggregation Function (ESAF) is used to obtainthe total sentiment scores for an given entity *ei*. We summed up the semantic orientation score of the opinion words in the sentence to determine the orientation of the opinion sentence. The *scores*(*s*) is normalized by the number of the opinion words, *n*, in the sentence to reflect the sentiment scores distributions of opinion words. The total sentiment scores of this entity will be aggregated by Entity Sentiment Aggregation Function. This score is normalized by the number of the sentences, *m*, and then the final sentiment score for an entity will ranges in the interval [+1, −1]. The sentiment are classified as SN (Strong Negative), N (Negative), Neu (Neutral), P (Positive) and SP (Strong Positive).

III. PROPOSED METHOD

The system performs an automatic sentiment analysis of Twitter messages about a product. The overall system canbe represented as shown in Figure 1. The data from web iscollected and stored. The parsing of data is done before analysis. Then we perform the actual sentiment analysis process on the stored data by retrieving it from database. And the final result generated is given as output to the users.

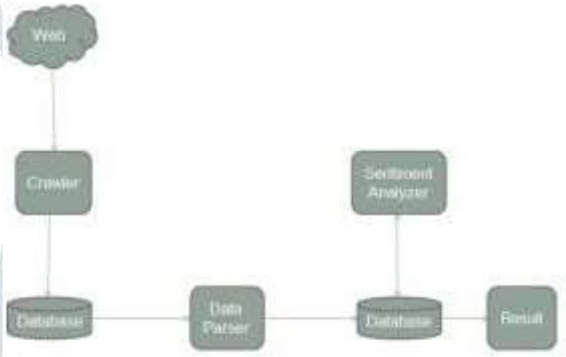

Figure 1: Overall System Architecture

The systems can be simplified and represented as shownin Figure 2.

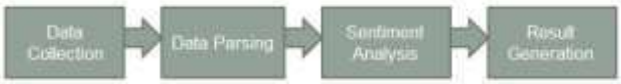

Figure 2: Simplified Block Diagram

The system can be divided mainly into following parts – the search part, the Sentiment Analysis part and the storage part.

A. SEARCH MODULE

In the searching part we need to enter the keyword to search. Based on the keyword entered a list of tweets are listed using crawling. Twitter API (Application Program Interface), which supports searching tweets pertaining to aspecific query is used to get the tweets. Searching is limited to a fixed number of tweets at a time. Content based focused web crawling is performed. The set of review tweets are stored using HDFS.

B. SENTIMENT ANALYSIS MODULE

Next part is the Sentiment Analysis. First the reviews are retrieved from database one by one. Then data parsing is applied to review tweets. Data parsing includes filtering process ie., removal of URLs, Twitter user names, Twitter special words, emoticons etc. from each tweets. Then tokenization is done which includes segmentation ofthe text into a bag of words. Then we remove stop-words from the text. Stop-words include articles like ‘a’, ‘the’ etc. Finally, we do n-gram construction from remaining text. The data parsing is done to remove all unwanted details from tweets and to obtain only the relevant parts for analysis. Now we need to compare the generated phrases or words with the set of words in dictionary. For that we need to construct a dictionary which consists of words that will be of use for analysis along with their synonyms and antonyms and a score for each term. The dictionary will be updated each time a new sentimentbearing term occurs in the tweet review during the analysis process. Now, we need to give score to each termin the review based on the given threshold values bycomparing with the entries in dictionary. Then we calculate the overall score of the review. After calculatingthe score, the results are generated in a structured format understandable by the user easily.

C. STORAGE MODULE

Storage has greater importance in our system as the system deals with large amount of data from web at a time referred as Big Data. Here we are using Hadoop for storage purpose. The Hadoop architecture consists of Hadoop Distributed File System, and a programming model, MapReduce, to perform data intensive computations on a cluster of commodity computers. The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware [7]. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets. Files in HDFS are write-once and have strictly one writer at any time. The architecture of HDFS is shown in Figure 3.

HDFS consists of two services namely, theNameNode and DataNode. The NameNode is acentralized, single server, responsible for maintaining themetadata for files inside HDFS. NameNode maintains thedirectory tree structure for the files in the filesystem. TheDataNodes store the files in the form of blocks on behalfof the client. Every block is stored as a separate file innode„s local filesystem. DataNode is responsible forstoring, retrieving and deleting blocks on the request ofNameNode. Files in HDFS are divided into blocks, with adefault block size of 64 MB, and each block is replicatedand stored in multiple DataNodes. The NameNodemaintains the metadata for each file stored into HDFS, inits main memory. This includes a mapping between storedfile names, the corresponding blocks of each file and theDataNodes that host these blocks. Hence, every request by a client to create, write, read or delete a file passthrough the NameNode. Using the metadata stored,NameNode has to direct every request from client to the appropriate set of DataNodes. The client then communicates directly with the DataNodes to perform fileoperations. HDFS can be accessed from applications inmany different ways. Natively, HDFS provides a JavaAPI for applications to use.

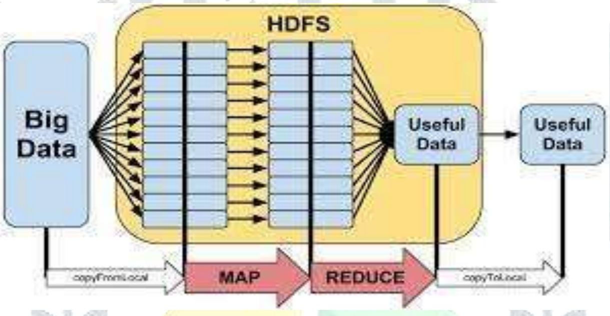

Figure 3: HDFS Architecture

The Hadoop uses the technique called MapReduce for efficient storage. It is the heartof Hadoop. The term MapReduce actually refers to two separate and distinct tasks that Hadoop programs perform.The first is the map job, which takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). The reduce job takes the output from a map as input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce job is always performed after the map job. Figure 4 showsthe MapReduce technique.

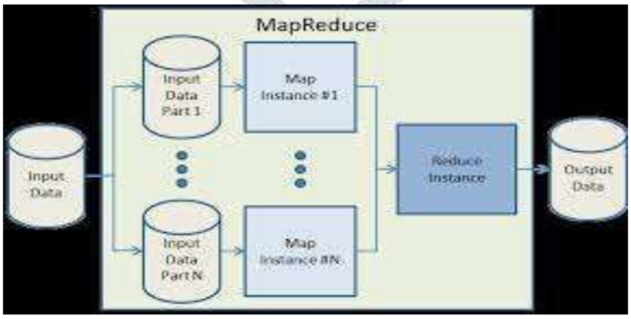

Figure 4: MapReduce Technique

IV. CONCLUSION

On deciding about the selection of a product, the best way is to rely upon the opinions of other users. With the advancement in web, user reviews can be obtained through various sources. Twitter is a major source of user reviews nowadays. There are several mechanisms proposed to analyze the sentiment from a review. Here thesystem analyses the sentiment hidden in a tweet about any specific product and provide the result to users in a structured format. The huge number of tweets produced, their storage and manipulation accounts to the Big Data and the problem arises regarding the efficient management of these vast data. We use HDFS for solving this problem. The previous methods related to product review mining and sentiment analysis are studied. Sentiment analysis is a difficult task and its difficulty increases with the complexity of opinions expressed. The system enables users to decide the most suitable product based on opinion of others. It transforms vast amount of data into an aggregated structured information. It providesa fast and less expensive alternative to traditional polls formining public opinions.

References

[1] Lada Banic, Ana Mihanovic, Marco Barkus, “Using Big Data and Sentiment Analysis in Product Evaluation.” In Proceedings of MIPRO 2013, May 20-24,2013.

[2] O. Medelyan, S. Schulz, J. Paetzold, M. Poprat, and K. Marko."Language specific and topic focused web crawling." In Proceedings of the Language Resources Conference LREC. 2006.

[3] W. Liu, Y. Hualiang, and X. Jiangu. "Automatically extracting user reviews from forum sites." Computers & Mathematics with Applications 62, no. 7 (2011): 2779-2792

[4] B.Pang, Lillian Lee, ShivakumarVaithyanathan, “Thumbs up? Sentiment Classification. using Machine Learning Technique”, In proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP), July 2001, pp, 79-86.

[5] Alexander Pak, Patrick Paroubek “Twitter as a Corpus for Sentiment Analysis and Opinion Mining**”,** In Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC\'10) (May 2010).

[6] X. Zhou, X. Tao, Z. Yang and J. yong “Sentiment Analysis on Tweets for Social Events”, In proceedings of the 2013 IEEE 17th International Conference on Computer Supported Cooperative Work in Design, pp, 557-561.

[7] K. Schvachko, H. Kuang, S. Radia, R. Chansler. “The Hadoop Distributed File System” In Proceedings of IEEE 26th symposium on Mass Storage Systems and Technologies (MSST), Incline Village, Nevada, USA, May 2010.