Analyzing Twitter Sentiments Through Big Data Analytics

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Engineering

in

Computer Science and Engineering

Submitted By

Monu Kumar

(Roll No. 801432013)

Under the supervision of:

Dr. Anju Bala

Assistant Professor

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

THAPAR UNIVERSITY

PATIALA – 147004

June 2016

Certificate

I hereby certify that the work which is being presented in the thesis entitled, **"*Analyzing Twitter Sentiments Through Big Data Analytics*"** in partial fulfillment of the requirements for the award of degree of Master of Engineering in ***Computer Science & Engineering*** submitted in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of ***Dr. Anju Bala*** and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

Monu Kumar
ME-CS
801432013

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

Dr. Anju Bala
Assistant Professor
CSED

Countersigned by

Dr. Maninder Singh
Head
Computer Science and Engineering Department
Thapar University
Patiala

Dr. S. S. Bhatia
Dean of Academic Affai
Thapar University
Patiala

Acknowledgement

First of all I would like to thank the Almighty, who has always guided me to work in the right direction. It is a privilege to thank my respected supervisor **Dr. Anju Bala,** Assistant Professor Computer Science & Engineering Department. She has been an cooperative guide and has been very supportive in achieving this task. This work would not have been possible without her guidance. I also thank my supervisor for her time, patience, discussions and valuable comments. Her enthusiasm and optimism made this journey rewarding. I am truly grateful to her for understanding whenever I needed help and correcting me whenever required. I am also heartily thankful to **Dr. Maninder Singh**, Associate Professor and Head, Computer Science & Engineering Department and Dr. Ashutosh Mishra, PG coordinator, for motivation and providing uncanny support throughout the preparation phase of this report.

I would like to express my gratitude to **Dr. S. S. Bhatia**, Senior Professor and Dean of Academic Affairs, for making provisions of infrastructure such as library facilities, computer labs equipped with net facilities, immensely useful for the learners to equip themselves with the latest in the field.

Last but not least, I would like to thank my family for their wonderful love and encouragement, without their blessings none of this would have been possible.

Monu Kumar
(**801432013)**

Abstract

The advent of social media has generated a lot of buzz among Internet users these days. A number of social networking sites are used these days, which has led to the rise of sentiment analysis. Twitter is a popular site where users post comments in the form of short status messages. Millions of tweets are received every year and sentiment analysis of these tweets interest among Internet users today. Data from these social networking sites can be used for a number of purposes, like prediction, marketing or sentiment analysis. Twitter is a highly used social media site for posting comments through short status messages. The millions of tweets received every year could be subjected to sentiment analysis. But handling such a huge amount of unstructured data is a tedious task to take up. The current Analytics tools and models used that are available in the market are not sufficient to manage big data.

In this thesis, we make use of Apache Mahout along with its Hadoop functionalities to carry out sentiment analysis Hadoop is a framework that performs computations over large datasets. With its framework MapReduce, it divides queries among different nodes nodes with computations to be performed in parallel. This provides faster query execution and faster result provision.

In this thesis, we take up the opinions of people on the services of Airtel. These opinions are converted into a training set and Mahout is used to carry out Naïve Bayes classification to decide how many tweets are correctly classified into being positive, negative and neutral.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 SENTIMENT ANALYSIS

When the product is bought for the first time, it needs to be chosen among various product having same characteristics. Organisations always advertise their product and brands by showing only positive points and characteristics of the product and hide the negative ones. So best way of selecting the product is to consider the opinions of the people who have already buy that product. By revising the reviews one can easily buy the best product. Analysis of such texts is an effective way to gather user information than the traditional structured data collection where people are usually not interested to give answers.

On the other hand, sentiment analysis listens to written reviews and answers in the form of positive and negative features of the product. Sentiment analysis or opinion mining is the process of knowing the people's attitude, opinions, their feelings and emotions about any product or movie. It is an information retrieval task as well as natural language processing task which is very difficult to perform but it is done due to its applications in many areas.

With world wide web(www) expanding its reach to anything and everything related to our daily lives, people are becoming more and more vocal to express their views and ideas on online portals, blogs etc. Whenever people buy the product, they want to know about that product. It can be known by the opinions of other people who have purchased that product. Most of the people share their good or bad experiences on review sites. These reviews help the consumers and business organization to get knowledge about that product. Millions of reviews are available on the web. Hence, it becomes very difficult for human beings to analyze the sentiments present in this huge amount of text. If there exists any method which able the computer to perform this analyses then it will be very beneficial for humans and can reduce their efforts. To extract sentiments about an object from this huge web, automated opinion mining system i.e. algorithmic method of analysis of large number of reviews is thus needed so that it can be easily and efficiently done by the computer.

We can say that sentiment analysis focus to know the attitude of the opinion holder in context of any product or to find the overall polarity of a sentence or document. Sentiment analysis can be of many type means polarity can be checked with different

methods. However, with extensive social media available on the web sentiment analysis is now treated as a big data task. Hence the conventional sentiment analysis approaches cannot effectively handle the large amount of sentiment data available these days. So, there is a need to perform sentiment analysis with effective methods which provides efficient results. To get effective results sentiment analysis is done by using MAHOUT tool. Brief description about MAHOUT tool is mentioned in next section.

1.2 MAHOUT

It is known as Apache Mahout which is an open source project. It is developed by ASF (Apache Software Foundation). Nowadays, machine learning algorithms are used in many application areas. The main objective is to create the scalable machine learning algorithms. Under Apache license, these algorithms are free to use. Mahout is a framework for data mining and generally it runs together with Hadoop system in its background. Hadoop can handle the large amount of data. Brief description about Hadoop is given in next section.

Machine learning algorithms like classification, recommendation and clustering are implemented using Mahout. It is chosen to implement the sentiment analysis due to its following features:-

- All the algorithms of the Mahout are implemented on the Hadoop, and it provides best platform to work in distributed computing environment. Apache Hadoop library is used by the Mahout so that it works well and effectively.
- Mahout provides readymade infrastructure or framework to coders for performing data mining tasks on huge amount of data.
- Mahout provides the application to analyze huge amount of data nery effectively and very fast.

1.3 HADOOP

Hadoop supports Java based programming framework. It is an open source framework and it works in distributed computing environment which forms clusters of computers. Apache software foundation sponsored the Hadoop. It helps in processing the large amount of data in distributed environment. With the help of Hadoop it becomes easy to run the applications on large systems having thousand

numbers of nodes and each contains thousands of terabyte. It provides very high data transfer rates due to its distributed file system. If there is failure of any node, system does not stop processing. This idea reduces the risk of terrible system failure even in the case, when large number of nodes gets failure. Basically, Hadoop is based on map reduce. It is a framework which breaks the application into smaller parts which are called blocks or fragments. These fragments can be run on different nodes. In general, Hadoop mainly consists of HDFS (Hadoop Distributed File System), map reduce, Hadoop kernel and some related projects such as Apache Hive etc. Google, IBM and Yahoo use Hadoop framework for the large applications like search engines and for advertising. For using Hadoop mainly Linux and Windows are preferred.

1.4 TYPES OF SENTIMENT ANALYSIS

Before doing opinion mining it is better to know that it can be done in many forms. Opinions may differ from one person to another person about the same product. Opinions may contain some positive and negative points about the same product because it may be possible that one product may have some good features and some poor features

On the basis of all these cases, sentiment analysis can be of many types. Sentiments are generally subjective information or expressions which describe people's feelings, sentiments or emotions towards an event or an entity. When sentiments are expressed in the form of tweets or reviews, instead of just saying 'Yes' or 'No', there is need to identify the actual sentiment or emotions that is there is a need of subjective analysis of each word expressed in review. Fig 1.1 shows the various levels at which the sentiment analysis can be done.

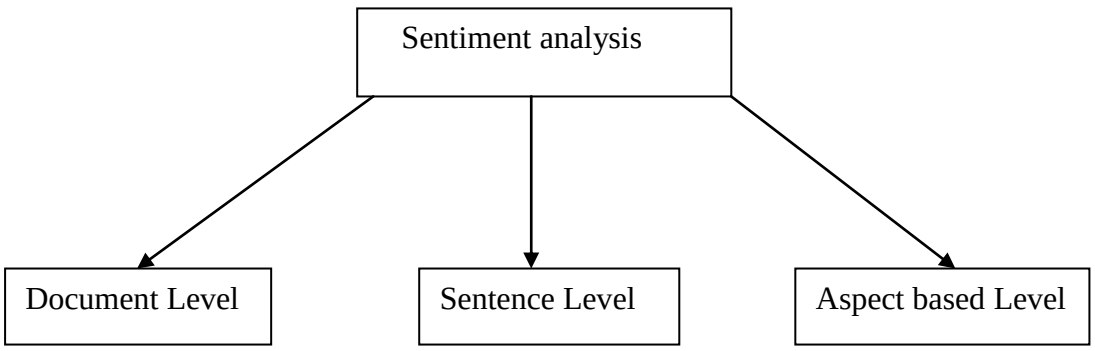


Fig 1.1 Types of sentiment analysis

1.4.1 Document level

In this type of sentiment analysis, every document mainly consist the opinions of the single opinion holder and about the single event or entity. And an opinion is classified into two classes either positive or negative. Mostly, each document expresses opinions like web review etc. But it is not the case that all the opinions belong to same document express the same views or sentiments about same entity. In this, each document contains different (some positive or some negative). But the result is given on the basis of whole document not on the individual opinion. Whole document is reduced to a single opinion score. Generally, document level sentiment analysis is done.

1.4.2 Sentence level

To get more refined view of different sentiments expressed in the document about the events or entities, there is a concept of sentence level. In this analysis those sentences are filtered out which contain no opinion and then determines whether the opinion on the entity is positive or negative. In this level of sentiment analysis a review or opinion may be subjective or opinion review. Subjective review contains the sentiment words so that it is easy to classify the review as positive or negative. For example:- "Few days ago, I purchased a phone. It is very good phone, its touch is very good. Its look is so great and clarity of voice is excellent. I like this phone."

So in this example by considering the words like good, great, excellent and like it can be easily identified that this subjective opinion is positive. So the opinions which contain subjective words or phrases so that polarity can be easily identified are categorised as subjective reviews.

Objective reviews are classified by the polls or stars. For example: movie ratings, song ratings are done in objective manner, sometimes rating of product is also is done in objective manner. If 4 to 5 stars are given, then it is considered as positive opinion and if 1 to 3 stars are given then it is considered as negative opinion. In this level of sentiment analysis, each opinion is considered and score is assigned to each opinion and overall result is found out by considering the score of each document and then it is analysed that whether the opinions are positive or negative.

1.4.3 Aspect based

In this level of sentiment analysis, it determines the polarity of opinions in which

review are expressed for different aspects or features of the same product. Entities may be mobile phone, laptop, camera etc. An aspect or feature is the component or an attribute of an entity. For ex: camera of a phone, battery life of phone, food quality of restaurant etc. The main advantage of this type of sentiment analysis is that there is a possibility to get the opinions about the interested feature of the product. Different features may have different sentiment responses. For ex: phone has good camera quality but screen is small.

Sentiment analysis is useful in many areas like in business, social media etc. and it has many advantages which are described in next section.

1.5 ADVANTAGES OF SENTIMENT ANALYSIS

- It provides benefits for the business purposes.
- It helps the people to know about the good or bad feature of the product they want to buy.
- It helps the organisation or company so that they can know the limitations or bad features of their product and hence can improve them.
- It is useful for the competitive party so that they can know the weakness of the product of opposition party and hence can launch the new product which overcomes the problem of that product.
- By sentiment analysis business organisation can know the success of their product and hence produce more products which got success.
- It reduces the effort of human being to analyse the product as good or bad.
- It is beneficial for social media analysis.

As there are many advantages of sentiment analysis, thus to perform sentiment analysis it is required to gather the opinions of different consumers. These can be gathered from various sources. Next section provides the various sources from where these opinions can be gathered.

1.6 SOURCES

There are mainly four sources are available to gather the opinions and these are:-

- ***Review sites***

  A huge and increasing body of people's reviews are available on internet. Mostly, opinions about products are expressed in unstructured manner. These opinions or

reviews can be gathered from the many e-commerce sites like www.yelp.com from where restaurants reviews can be collected, www.amazon .com from where reviews about product can be collected and www.reviewcenter.com on which million of reviews of consumers about product are available. Many other professional sites are also available like www.zdnet.com and consumer sites for products are www.epinions.com, www.consumerreview.com.

- ***Blogs***

As, nowadays many people are using internet due to which blog pages and blogging are also growing rapidly. Bloggers keep recording the daily events of their lives and share their feelings, opinions and feelings in form of a blog. Mostly, these blogs contains the reviews about many issues, products etc.

- ***Micro-blogging***

Nowadays, most popular micro-blogging service provider is twitter, where people express the opinions in the form of messages called as tweets. These reviews express the reviews about different entities or event. Most of the twitter messages are used in the form of data sources to classify the sentiments.

Initially, it was easy to access the tweets but now to increase the security from the unauthorised users it is not easy to access the reviews from twitter or tweets. Because twitter does not allow unauthorised users to access the tweets. So to access the tweets, many steps have to follow like first of all there is a need to generate an API (App Program Interface). After creating an API twitter provides an authorised key and consumer key. By using or having these keys an user can access the tweets from the twitter. But there is one condition that to create an API, the account on tweeter must be created at least three months ago, otherwise your account will not be able to access the tweets.

- ***Data sets***

Mostly, sentiment analysis is done on movie reviews and readymade dataset is used for classification. Especially for movie reviews, dataset is available on www.cs.cornell.edu. Not only for movies, in fact datasets are also available as MDS (Multi Domain Set) dataset. These can be accessed from www.cs.jhu.edu. These MDS dataset mainly consists of opinions about four different types of products. And these product reviews are derived from Amazon .com which includes the reviews about books, kitchen and electronic appliances and DVDs.

Whereas these contains 1000 negative and 1000 positive reviews for each product. Many other sites are also available for review datasets like www.cs.uic.edu.In this, dataset contains the reviews about five electronic products which are extracted from Cnet and Amazon. Now, by accessing the reviews from any one of the source, sentiment analysis is done on these reviews. Among all the sources it is most difficult to access the reviews from twitter and most convenient and easy to use the datasets because in this there is no effort is done and mostly equal number of positive and negative reviews are accessed. So we can analyse the results more accurately.

After collecting the reviews, process of sentiment analysis get starts. There are many techniques and methods are available to classify the polarity of sentiments. All the techniques and methods are described in next section. Basic technique for the classification of reviews is to use classifiers which are of many types. Classifiers are used to classify the sentiments in two different classes that is one positive and other is negative. In which class, a review falls, it depends on its polarity which is found out by the words or sentiments used in an opinion.

1.7 METHODS FOR SENTIMENT ANALYSIS

Broadly, methods for sentiment analysis can be divided into two main categories as shown in fig 1.2.

- **Supervised Machine Learning**

  This technique mainly contains Support vector machine, Naive Maximum Entropy methods. These are classifiers and classify the opinions in two classes. These classifiers need the training set data and test set. Based on these inputs, these methods classify the reviews.

- **Unsupervised Machine Learning**

  This technique contains clustering and lexical based analysis and it does not need training set data. Results of polarity of sentiments may vary from one technique or method to another technique and accuracy may differ.

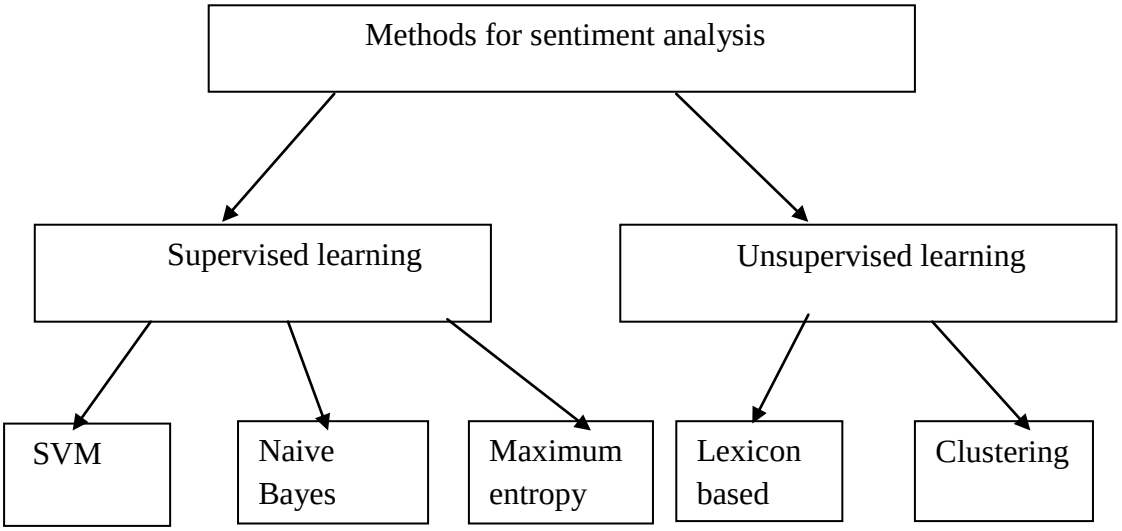


Fig 1.2 Methods for sentiment analysis

Now, the methods of supervised learning are described.

1.7.1 Support vector machine

In support vector machine, each data point is considered as k-dimensional vector which is a list of k numbers. It is checked that whether these points can be seperated with k-1 dimensional hyperplane. It is known as linear classifier. Data is classified by these hyperplanes.

1.7.2 Naive bayes

Naive Bayes gave an effective method to carry out the study of classification. It works on the following principles. Suppose there are n possible classes $X=\{x_1,x_2\ldots\ldots x_m\}$ for a domain of documents $Y=\{y_1,y_2\ldots..y_n\}$. Let $W=\{w_1,w_2\ldots w_n\}$ be a set of words that appear in the documents of Y.

Then

$$P(x|y) = \frac{P(x)P(y|x)}{P(y)} \qquad \ldots\ldots\ldots\ldots(1)$$

Naive Bayes assumed that each word in W appears independently. Thus, this formula takes the form as shown below:

$$P(x|y) = P(x)\prod_{n=i}^{i_d} P(w_k\ |x)_k^t \qquad \ldots\ldots\ldots\ldots(2)$$

where i_d is the number of unique words in document d, t_k is the frequency of each word w_k.

By applying Naive Bayes classifier, we can estimate P(x) and $P(w_k|x)$ as:

$$\hat{P}(x) = N_x / N$$

.....(3)

$$\hat{P}(w_k|x) = \frac{N_{wk}}{\sum_{wiew} N_{wi}}$$ .....(4)

N is the total number of documents, N_x is the number of documents in class x. N_{wi} is the frequency of word w_i in class x.

1.7.3 Maximum entropy

This classifier is a probabilistic classifier but it does not assume features to be conditionally independent of each other as naive bayes assumes. It is based on the principle of maximum entropy and chooses the training model with the largest entopy. There is binary feature indicator is used which has value 0 if the document does not belong to the class and gives the output as 1 when particular document belongs to class c and it contains the word w. The statistics of training dataset is expressed as the expected value of most appropriate binary valued indicator function. It uses empirical probability distribution to create a model, assigning terms by considering contextual information.

Now, unsupervised learning methods are described.

1.7.4 Lexical based

In lexicon based unsupervised learning method, there is a sentiment dictionary and it doesnot require to store large amount of data corpus and training set. There are many lexical resources like sentiword net, sentic net, MPQA etc. These contain the three numerical score for each term as positive score, negative score and objectivity score (neutral). Total score of a term is calculated as positive score minus negative score. If total score is positive then term is considered as positive otherwise as negative feature.

1.7.5 Clustering

In clustering technique, terms having similar features belong to one cluster and so there are two clusters are created. One cluster is for the opinions having positive

feature and another for the opinions having negative features. And neutral opinions do not belong to any of the cluster. In sentiment analysis, k-means algorithm is used for clustering. In clustering, there are centroids and for each term distance is calculated from all the centroids and term belongs to that cluster from which its distance is minimum.

1.7.5.1 Fuzzy c-means clustering

Fuzzy c-means clustering comprises of two processes: in the first phase cluster is calculated and in the next phase, points are assigned to centres by using a formula of distance. This is done till cluster centres get stabilized. The algorithm similar to k-means clustering in many ways but it gives data items a membership value between 0 to 1. Thus it includes a concept of partial membership and provides overlapping clusters to support it. A fuzzification parameter m in the range [1,n] is needed to determine the degree of fuzziness in clusters. When m tends to 1, the algorithm functions like a crisp partitioning algorithm . For larger values of m the overlapping in clusters becomes more. The algorithm calculates the membership value μ with the formula,

$$\mu_j(x_i) = \frac{\left(\frac{1}{d_{ji}}\right)^{\frac{1}{m-1}}}{\sum_{k=1}^{p}\left(\frac{1}{d_{ki}}\right)^{\frac{1}{m-1}}} \qquad \text{.....(5)}$$

where

μj(xi) : membership of x_i in cluster j

dji : distance of x_i from cluster c_j

m : fuzzification parameter

p : number of clusters

d_{ki} : distance of x_i in cluster C_k

The new cluster centers are calculated with these membership values

$$c_j = \frac{\sum_i \left[\mu_j(x_i)\right]^m x_i}{\sum_i \left[\mu_j(x_i)\right]^m} \qquad \text{.....(6)}$$

where Cj : is the center of the jth cluster

xi : is the ith data point

μj : the function which returns the membership

m : is the fuzzification parameter

This is a special form of weighted average. Modify the degree of fuzziness in xi's current membership and multiply this by xi. The product obtained is divided by the sum of the fuzzified membership. This way new centroids are calculated for clusters. And this process is stopped when values of successive centroids become same. And on the final centroid is determined and clusters are formed on the basis of that clustering. Finally, two clusters are formed one for positive featured terms and another for negative featured terms. So to get more accurate results and to make efficient classification there is a need to remove the data from reviews which is not required and this process is called pre-processing. In this, unnecessary words like name of opinion holder, date of review etc. are removed. Pre-processing phase is described in next section and it is very important for sentiment analysis.

1.8 PREPROCESSING

So, to remove unnecessary words which are not required for sentiment analysis, preprocessing of data is done which includes following steps which are described in fig 1.

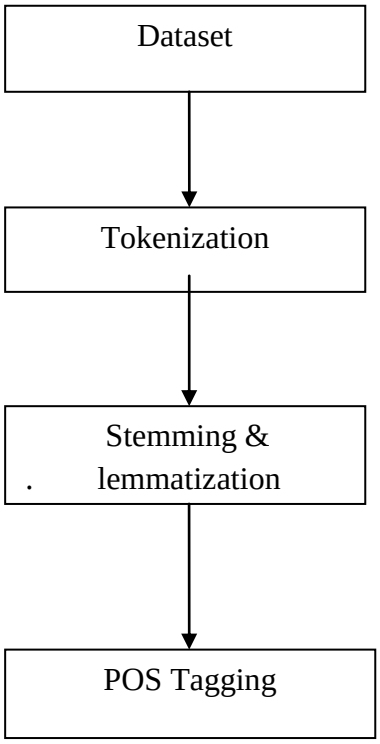


Fig 1.3 Pre-processing

So, initially dataset is collected and then tokenization is done and output of the tokenization is terms. On these terms stemming and lemmatization is done and after that POS tagging is done. After all these steps pre-processing is finished.

Now, all the steps are described in detail:-

1.8.1 Dataset

Initially dataset is created by collecting data from social media and other sources.

1.8.2 Tokenization

It is a process of breaking the sentence into its constituent parts. And there is not only one process or method to do tokenization. Different algorithms are available for this for different apps. It is very important in the field of sentiment analysis.

- ***White space tokenizer***

The whitespace tokenizer simply breaks the sentences and divides the text on any sequence of whitespace, tab, or newline characters. For example: "I saw a movie, it is nice and songs are very good". After applying white space tokenizer output is I, saw, a, movie, it, is, nice, and, songs, are, very, good.

- ***Additional Punctuation***

Punctuation must be considered at the tokenization stage. So, the goal for tokenizing is to properly differentiate between various senses for individual punctuation marks. The basic strategy for handling punctuation is by trying to identify all the word-internal marks or symbols first, so all others can be tokenized as separate elements. Some of these considerations:

- Generally, words can contain combination of numbers, letters, hyphens, apostrophes and underscores.
- Sometimes words may contain digit and comma and must be tokenized as one word.

The rest of the punctuation can be divided as separate words. Generally, this means exclamation points, question marks, and dollar signs without any subsequent digits. It is obtained that it works perfect to tokenize the sequences like !!! into three exclamation marks separately.

1. If tokenizer is good then it will provide good results especially if amount for training data is limited.
    - It increases the effectiveness of classifier.
    - Increase the portability of the model.
2. There is need of performing careful tokenization if huge volume of data is present.
3. In sentiment analysis, good tokenizer is very important because reviews contain a lot of punctuation and these are present in unstructured form.

1.8.3 Stemming and lemmatization

Due to grammatical reasons, sentences may contain the different forms of a word like decorate, decorates and decorating. In addition to these, there are many derivationally related word which possess the similar meaning like automata, automatic and automatically. While doing sentiment analysis, these types of words can be analysed as same word. So the aim of both lemmatization and stemming is to reduce these types of forms and reduce the word from its derivational form to its corresponding base form.

For example: am, are, is $\Rightarrow$ be

girl, girls, girl's, girls' $\Rightarrow$ girl

The result of this mapping of text will be something like:

the girl's cars are different colors $\Rightarrow$ the girl car be differ color,

whereas, the two words are used in different senses.

Basically, stemming is a process of chopping the suffix of words and also it removes the derivational affixes.

Generally, lemmatization is a process of doing things in a proper way by using vocabulary and morphological analysis of words, it aims to remove only the extensive endings and convert into dictionary or base form of the word and it is called as lemma. Let us consider a token as saw then after stemming the output will be only s, whereas after lemmatization output will be either saw or see. It depends on whether the token is used as verb or noun.

Lemmatization and stemming are different from each other as stemming generally reduces the derivationally related words and lemmatization generally reduces the different forms of basic words or the extensions to the root word. Most effective algorithm for stemming is Porter's algorithm. It contains some rules which are used for stemming. From the following rule, one rule is selected for each term and rule is selected which has longest suffix.

EXAMPLE	RULE		
IES →	I	Ponies →	Poni
SSES →	SS	Dresses →	Dress
S →		Chapattis →	Chapatti
SS →	SS	Caress →	Caress

1.8.4 Pos tagging

POS means Part Of Speech: Generally, POS tagging is a process based on sentences, every sentence is formed using sequence of words and POS tagging is used to label or tag each word with its corresponding part of speech. But sometimes, there might be ambiguity between words because most of the words in the language may have more than one part of speech. Or sometimes there are many words in the sentence for which the tagger has no knowledge about those words. So the solution of these problems is to use the context around the target word and choose the most probable tag for that word by using information provided by the context words.

Part Of Speech Tagging also called as grammatical tagging or word category disambiguation, it is the process of labelling a word in a text with respect to its part of speech, on the basis of definition and its context. Context means its relationsh with adjacent as well as related words in a sentence, phrase or paragraph. A simplified form of this is the word identification as adjectives, nouns, adverbs etc.

For example: "The sailor dogs the hatch."

Here, dogs which is generally considered as plural noun but in this sentence it has to be considered as verb. So in this case, grammatical context is the way to determine this as a verb. In schools, there are generally nine parts of speech – noun, adjective, verb, adverb, interjection, conjunction, pronoun, article and preposition. But in addition to these there are many part of speech like NN for singular common nouns, NNS for plural common nouns, NP for singular proper noun nad so on. So, to handle

all these part of speech tagging stochastic method for tagging is used which uses almost one thousand parts of speech. So that computer can consider every case of pos tagging.

After performing all these steps, pre-processing phase is finished. After this, methods of sentiment analysis is applied. As we disscussed about the various methods of the sentiment analysis in previous chapter. Now, we will describe the work done related to these methods. Many researchers worked on supervised learning methods, lexical based and clustering methods and details of their work done is mentioned in next chapter.

Chapter 2: Literature Survey

2.1 METHODS TO PERFORM SENTIMENT ANALYSIS

As, we have discussed in previous chapter about the sentiment analysis, its types and methods for sentiment analysis. Many methods are available for sentiment analysis and different methods provide different accuracy in results. Sentiment analysis is very useful for customers and for business organisation. It also provides the reason of success or failure of the product. For example "Moto G does not have better camera quality". By considering the reason of disliking the product, business organisation can improve the negative points of their product. But it is very challenging for computer to do sentiment analysis because the reviews are not directly understood by the computer and these contains noise which needs to be separated out before performing sentiment analysis.

Early Lee and Pang [1] applied different methods to identify the polarity of product reviews and movie reviews respectively. They worked at document level. They used the support vector machine and naive bayes as classifiers and find the polarity as positive and negative. But these classifiers are not directly applied on the dataset which are collected from the reviews because reviews are in unstructured format. So if we directly apply these methods on dataset then it will not provide accurate results. Because initially reviews may contain misspelled words or noisy data which are not required for sentiment analysis and hence increase the overhead. To remove this noise, pre-processing is done in efficient manner which has described in previous chapter.

After pre-processing phase, feature extraction is done. In feature extraction, important terms are extracted. Terms which show the sentiments of the opinion holders are considered as important for sentiment analysis. There are various methods for feature extraction and accordingly provides different results. Many authors provided the different methods for sentiment analysis by using different feature extraction method. As it was discussed in previous chapter that sentiment analysis can be done by various methods.

A lot of work has been done in sentiment analysis by using different methods and their corresponding results are also provided. In next section, all the work done

related to opinion mining is described. Review is done on the basis of different technology as supervised methods, lexicon based and clustering.

2.2 SUPERVISED LEARNING METHODS

These techniques make use of a training and testing set for classification. Input feature vectors and the respective class labels form the training set. Training set classifies input feature vectors into respective class labels. Then a test set tests the model by generating the class labels from feature vectors not yet seen. Many machine learning methods like SVM (Support Vector Machine), ME (Maximum Entropy) and NB (Naive Bayes) are used for classification of reviews [2]. Some of these features that are used for opinion mining are negation, Term frequency, Term presence, Part-of-Speech and n-grams [3]. These features determine polarity of words, sentences and documents. Polarity could be negative, positive or neutral.

The best classification method for the text is SVM (Support vector machines) and it is considered as discriminative classifier [4]. According to the computational learning concept, a statistical classification method called support vector machine is based on the structural risk minimization principle. Support vector machine finds the decision surfaces which help to classify the training set into two classes as negative and positive. On the basis of support vectors decisions are made. Support vectors are the effective data points among the training set. There are many variants of support vector machine are developed. Among these variants Multi class support vector machine is mostly used for opinion mining.

Cui [5] proposed that most appropriate classification for sentiment analysis is Support vector machine because it can provide better results when opinions consist of both negative and positive terms. But when size of training data set is very small, in that case Naive Bayes classifier provides appropriate results. Because Support vector machine is a classifier of high quality and to develop this quality of classifier, Support vector machine needs a large set of training data set.

Domingos [6] concluded that Naive Bayes provides efficient results in case of certain problems where features are highly dependent. Whereas, Naive Bayes has basic assumption that features must be independent to each other but Naive Bayes gives better results.

Zhen Niu [7] proposed a new model to increase the efficiency of Naive Bayes. In his model, most efficient methods for computing the weight, classification and feature extraction are applied. By using the term frequency weight is calculated. Bayesian algorithm is used for this new model. In this, unique feature and representative features are used to adjust the weights of the classifiers. The information which represents a class is known as Representative feature. The information which helps in differentiating between the classes is known as Unique feature. On the basis of weights, probability for each classification is calculated and this feature improves the Bayesian algorithm.

Pak [8] generated a large amount of twitter data by collecting the tweets automatically. Tweets are collected with the help of API (Application Programming Interface). He analysed them by using emoticons. POS-tags and N-gram features are used by Naive Bayes classifier. But in this method, there is high probability of error because sentiments of tweets in training data set considered as only on the basis of polarization of emoticons. Reason for less efficiency of training set is that it considers only those tweets which have emoticons.

Xia et al. [9] proposed an ensemble architecture for sentiment classification. Ensemble framework is generated with the combination of classification methods and feature sets. In their approach, three base classifiers and two kinds of feature sets are used to create an ensemble framework. These two forms of feature sets are generated by using Word relations and Part of speech information. The three base classifiers are chosen as Maximum Entropy, Naive Bayes and Support Vector Machine. They used different ensemble techniques like weighted combination, fixed combination for opinion mining and accuracy was better as compared to previous one.

Neethu and Raajsree [10] generate ensemble classifier in their research. They did pre-processing then feature extraction is done in two phases. First, twitter specific features are retrieved. To find out the sentiments in the tweets emoticons and hashtags are used. Emoticons can be positive or negative according to which different weights are assigned. Emoticons which are positive are assigned a weight of ’1’ while negative ones are assigned’-1’. There could be negative as well as positive hashtags. Therefore the number of positive negative hashtags are taken as two different features in the feature vector.

Thus feature vector is comprises of 8 relevant features. These features are a part of speech (pos) tag, presence of negation, number of positive and negative keywords,

emoticons, number of positive and negative hash tags. The base classifiers used are Naive Bayes, Maximum entropy and SVM. Here a voting rule is used to create an ensemble classifier. The classifier will do classification on the basis of output of majority of classifiers. Using Twitter API, product related tweets are gathered. A dataset is made by gathering 1200 twitter posts related to electronic products. Dataset is split such that there are 1000 tweets in the training set and 200 in the test set .They used Stanford postagger1 for retrieving a POS tag from the collected tweets. Since a product domain is selected, subjective and objective tweets need not be analyzed separately. Both qualities contribute similarly to identify the product's quality.

This shows that domain information affects sentiment analysis. They used three types of basic classifiers (SVM, Nave Bayes, Maximum Entropy) and ensemble classifier for the purpose of classification of sentiments. Similar performance is given by these classifiers. As compared to others, Naive Bayes has better precision but accuracy and recall are lower. SVM, Maximum Entropy Classifier and Ensemble classifiers are similar in these evaluation metrics. While there accuracy is 90% whereas Naive Bayes has the accuracy of 89.5%. This shows the quality feature vector chosen for product domain. This feature vector helps in improved sentiment analysis despite of the classifier used.

It was all about the supervised methods but lexical based techniques are also available for sentiment analysis. These lexical based techniques are also known as symbolic techniques. All the work done related to these techniques is mentioned in the next section.

2.3 LEXICAL BASED METHOD

In lexicon based unsupervised learning method, there is a concept of sentiment dictionary. There are various lexical resources available for sentiment analysis and these are described in this section.

Senti word net is one of the lexical resources for sentiment analysis. it provides the three numerical sentiment score as negative, positive and neutral for each synset of WordNet. Synset means synonyms of the words or words which are related to each other like automobile and car. So, it provides synset based representation means for each synset there is different scores are assigned so in the situation where same term has different senses then it will give different sentiment scores. So to get the most

accurate meaning of the word, it needs to be coupled with WSD (Word Sense Disambiguation) algorithm.

Kamps et al. [11] performed the sentiment analysis by using WordNet as lexical database [12]. It is used to find out the sentiment words from a review with different dimensions. They created a distance metric on the basis of WordNet and find out the polarity of the adjectives. A lexical resource WordNet database includes the words which are related to each other by synonym relations.

Turney [13] did sentiment analysis by using bag of words. To calculate the overall polarity of all the sentiments, polarity of each term or word in a documrnt is calculated. Some aggregation functions are applied on these values. He determined the polarity of an opinion on the basis of average polarity of words which are extracted from the opinions. Those words are considered which are adverbs or adjectives.

It is lexical resource which is used for sentiment analysis at conceptual level. It is different from all discussed resources because it includes some complex or advanced concepts like to complete the goal and to celebrate special occasion etc. Now, sentic net assigns the sentiment scores between the range -1 and 1. These scores are assigned to approximate 14,000 common sense concepts. Sentiment score to each term is assigned on the basis of sixteen basic emotions. These emotions are defined in a model known as hourglass of emotions.

Many researchers worked on sentiment analysis using sentic net or the combination of word net and sentic net which is described in this section.

Yun quing xia [14] used sentic net and Bayesian model for contextual concept polarity disambiguation. Mauro [15] proposed a merged framework that merges word net, concept net and sentic net to extract key concepts from a sentence.

A system was proposed [16] that lets create their own sentiment analysis framework. In their framework, they combine sentic net, sentiword net and other sentiment analysis method.

Another proposed work [17] used senti net to extract bag of concepts and polarity features for subjectivity detection and other sentiment analysis task. Jay kuan [18] used sentic net concepts as seeds, giving a concept of random walk to obtain more concepts and polarity scores. Other proposed have proposed the joint use of knowledge bases and machine learning for twitter sentiment analysis, [19] short text message classification [20] and frame based opinion mining [21]

Erik cambria [22] proposed that how sentiment analysis is done using hybrid approach that by using machine learning and sentic net. In his approach, if opinions containing the words which are not found in the senti net then these opinions are classified by deep learning methods like support vector machine, naive bayes etc. instead of ignoring these opinions and hence it provides better results as compared to only lexical based methods.

Before applying sentic net, reviews are passed through dependency tree which provides the output as related words and then it is passed through semantic parser.

Now, there is another technique for sentiment analysis that is clustering. The work done related to clustering is presented in next section.

2.4 CLUSTERING METHODS

The process of collecting the objects of similar characteristics into one group is called clustering. Clustering is done at document level. Document clustering form topics by grouping documents without any idea of the category structure that exist in the collection. He documents are used to obtain subjective information. The main objective of clustering is to determine the structure in data to form the groups. Mostly, k-means algorithm is used for sentiment analysis. In this, centroid for each cluster is found. Similar terms are grouped into one cluster.

As k-means means algorithm [23] makes use of the vector space model by iteratively optimising k centroid vectors representing clusters. These clusters are updated by making use of the mean of nearest neighbours of the centroid. The algorithm proceeds to iteratively optimise the sum of squared distances between the set of vectors and their nearest neighbour centroid. This can be done by iteratively updating centroid and again assigning nearest neighbours to obtain new clusters till centroid values are the same.

A better model on opinions of mobile phones is presented in [24]. Synonym feature words are created by using clustering and support vector machine classification. In this, feature words are extracted by making use of adjective words. Senti word net assigns the polarity of the words along with their weight. And after that k means clustering is applied and results are achieved.

An unsupervised clustering technique proposed in [25] to categorise the product features as negative or positive. They used three kinds of context information these

are context information, opinion words and group information of opinion words. They used k-means and it was found that it gives good results by considering only opinion words rather than full word context.

Sentiment based clustering by lowering the number of object features [26] make use of the review data from web and summarizes review sentences of the same product. An information system is created to manage the implicit features. As there are high feature dimensions, feature dimension reduction algorithm is adopted. These retrieve the product opinion along feature and assign a rank to each aspect of product. They used k-means clustering technique and gave a rank to each facet of the product. Proposed method evaluated the effective clustering results.

[27] gave a technique for feature level opinion mining. Opinion words are used to gather the implicit and explicit features. Feature clusters are obtained by using k-means on the basis of three aspects: related opinion words, similarity and structure of features. Remove low frequency features. Finally clusters represented the polarity of objects.

Gang li and Fie Liu [28] proposed a clustering based approach using k-means clustering technique. Data is pre-processed by using stemming. Then TF-IDF weighting method is used to identify the weight of each term. Weight is calculated by the term frequency matrix. By considering TF-IDF accuracy was increased. But there was low accuracy and the lowest was 59%. But the results were unstable due to random selection of centroid in K-means algorithm. To get stable results voting mechanism was used. K-means clustering method is used many times. Result of each clustering process is considered as vote. A document is considered as positive if it contains more than half positive votes, otherwise it will be considered as negative. By using this performance become better. Lowest it was 75.17% and highest was 78%.

Till now, we have mentioned all the work done related to sentiment analysis and available methods for analysing the reviews and sometimes combination of methods are also used to improve the accuracy of results. As we discussed hybrid approach in which sentiment analysis is done by machine learning and lexicon based method. But still there are some problems with existing methods. These are described in next chapter.

Chapter 3: Problem Statement and Objective

3.1 PROBLEM STATEMENT

Social media has gained a lot of light in the past few years. Twitter has become an important platform which people are taking up to express their views and opinions about any topic. This microblogging service witnessed as large as 465 million accounts in the year 2012, which generated 175 million tweets per day. The number is exponentially rising with the growing popularity of this website. Sentiment analysis has made it possible to analyse the moods of a person. It can help us to decide the positive, negative or neutral views of a person based on his attitude on a given topic. Previously, it was used for lexical or syntax feature extraction, assigning a polarity label to each given document. These days, sites like Twitter show the influence that surroundings have on online users. It is tough to process big data using traditional techniques. The current Analytics tools and models used that are available in the market are not sufficient to manage big data. Therefore, there is a need to use a cloud storage for such type of applications. So we have utilized Hadoop for intelligent analysis and storage of big data.

To analyse such enormous amount of data, we make use of Apache Hadoop and its functionalities on the Mahout framework. Hadoop is a framework that performs computations over large datasets. Cloud computing like Hadoop helps to to perform operations on distributed data in an efficient manner. Hadoop has an internal framework called MapReduce to perform its functionality. Queries are divided among different nodes, to be performed in parallel. This is known as the Map stage. Then results are combined in the reduce stage to give an output. This provides faster query execution and faster result provision.

In this thesis, sentiment analysis is done on people's opinions of Airtel. Two techniques have been used:

i) Naïve bayes classification
ii) Fuzzy c means clustering

People have expressed their satisfaction or dissatisfaction with regard to this service provider, which in turn are used to analyse the number of people who are satisfied with the use of Airtel, as well as their judgement on its service. Mahout commands are used to perform Naïve Bayes classification and fuzzy c-means clustering which

helps to accurately classify tweets into their respective classes in this form of supervised learning.

3.2 OBJECTIVES

The main objectives of sentiment analysis performed on tweets of users with the help of Apache Mahout are:

- Hadoop is used to carry out sentiment analysis on Twitter, and it is incorporated the use with its machine learning tool **Mahout** which is a simple, scalable, research oriented workbench.
- Analysis was done using Apache Mahout (for data retrieval, preparation, and computation) and Excel (for plotting) on 80,000 tweets, where 70% of the dataset was taken as the training set.
- Naïve Bayes classification and fuzzy clustering is performed on the testing set to discover how many instances are positive, negative and neutral.
- It is found that 20% of the tweets are negative, 38 % are neutral and 40% are positive. Fuzzy clustering clubs the top terms together and sentiment analysis is done based on the similarity between them.

Chapter 4: Proposed Methodology and Tools Used

4.1 ABOUT APACHE MAHOUT

The growing success of companies and the rage of internet has made it essential to perform analysis and operations on a huge amount of data. Such tools can effectively enhance and organize data which needs processing thousands of messages or terabytes of content. Therein the need to use machine learning techniques has risen be it clustering, classification or generating recommendations .

Apache Mahout is one such intelligent application used to learn data from users. It helps to carry out machine learning operations in an effective and scalable manner. Mahout is thus a popular tool used to cluster documents, classify tags, organize content and perform collaborative filtering to give recommendations.

Currently, these tasks have been implemented using Mahout in areas that use real life applications:

- Collaborative filtering
- Clustering
- Classification

Mahout makes use of Apache Hadoop when it comes to implementation on huge datasets. Hadoop is an open-source framework that is used to store and perform operations on big data in a distributed environment by involving clusters of computers. It is capable of scaling up from single servers to thousands of machines, with each providing local computation and storage. It makes use of a framework called MapReduce which provides a well defined API for parallel computation tasks.

Mahout has been used to implement the following tasks:

- MapReduce enabled clustering implementations like K-means, Fuzzy K-means, Canopy and Dirichlet.
- Naive Bayes and Distributed Naive Bayes classification
- Distributed fitness function capabilities
- A Taste Project for implementing Collaborative filtering used in generating recommendations.
- Matrix and vector libraries.

4.1.1 WORKING WITH MAHOUT IN UBUNTU

In this thesis, we have worked with Mahout in Ubuntu. Ubuntu is an operating system based on Linux which runs on many devices like PCs, smartphones etc. It is open source and has the capability to support cloud and big data operations. Working with Mahout requires the following prerequisites:

- Install JDK 1.6 or higher.

- Add a dedicated user account for running Hadoop by using the following commands:

  $ sudo addgroup hadoop

  $ sudo adduser --ingroup hadoop hduser

- Install SSH so that Hadoop can suppot remote machines.

  user@ubuntu:~$ su - hduser

  hduser@ubuntu:~$ ssh-keygen -t rsa -P ""

  RSA is an encryption method used to create a key with a password. Once this is done, we can easily connect our local machine with the new Hadoop user created.

- Download Hadoop from one of the Apache Mirrors and extract contents to package /usr/local/.

  cd /usr/local

  $ sudo tar hadoop2.4.1.tar.gz

  $ sudo mv hadoop-2.4.1 hadoop

- Update environment variable in files :env.sh,core-site.xml, mapred-site.xml and hdfs-site.xml

- Start Hadoop by using the following command:

  /usr/local/hadoop/bin/start-all.sh

  The namenode,datanode, taskTracker,jobTracker get activated using this command.

- To install Mahout, install Maven 2.0.10 or by using the following command now:

  $sudo apt-get install maven

4.2 METHODOLOGY I: CLASSIFYING TWEETS USING NAÏVE BAYES CLASSIFICATION

Mahout can be used to classify content based on Bayesian statistics. Naive Bayes classifier is an efficient and accurate classifier. A MapReduce based Naive Bayes is used for this purpose, to compute results on huge datasets in a parallel manner. This classifier might not work very well when training data is not balanced. The second approach implemented in Mahout is the Complementary Naive Bayes classifier which deals with problems in the basic approach. In this thesis, MapReduce based Naive Bayes classifier has been used to perform sentiment analysis on the gathered tweets.

Using Naive Bayes classifier is a two-part process as it keeps in view the features or words associated with a document and then this information is used to categorize or classify unseen content. The first step creates a training set which creates a model by looking at the content already classified, which in our case is categorized by labels positive, negative and

In the next step, the testing set containing the unseen content is classified by using this classifier. Thus the tweets will be predicted in a particular category i.e. positive, negative or neutral.

Before running the classifier, a training set is created by creating a set of documents with a unique id. These documents are placed in labelled folders (positive, negative and neutral) . The tweets in these documents act as values. Thus sequence files can be created in Mahout using this <id,value> pair. The training data is preprocessed before going into the classification phase.

The tweets are collected over a period of two months, September and October. A Python is linked to Twitter API and the authentication details are used to capture the required tweets. The main phases of tweet collection are:

1. **Data Collection** - Our code is integrated with the Twitter API v1.1 using consumer keys/secrets and access token key/secrets. Finally, we prepare a

training set by creating three folders- positive, negative, neutral. Tweets are stored as values in files and names of documents are tweet ids.

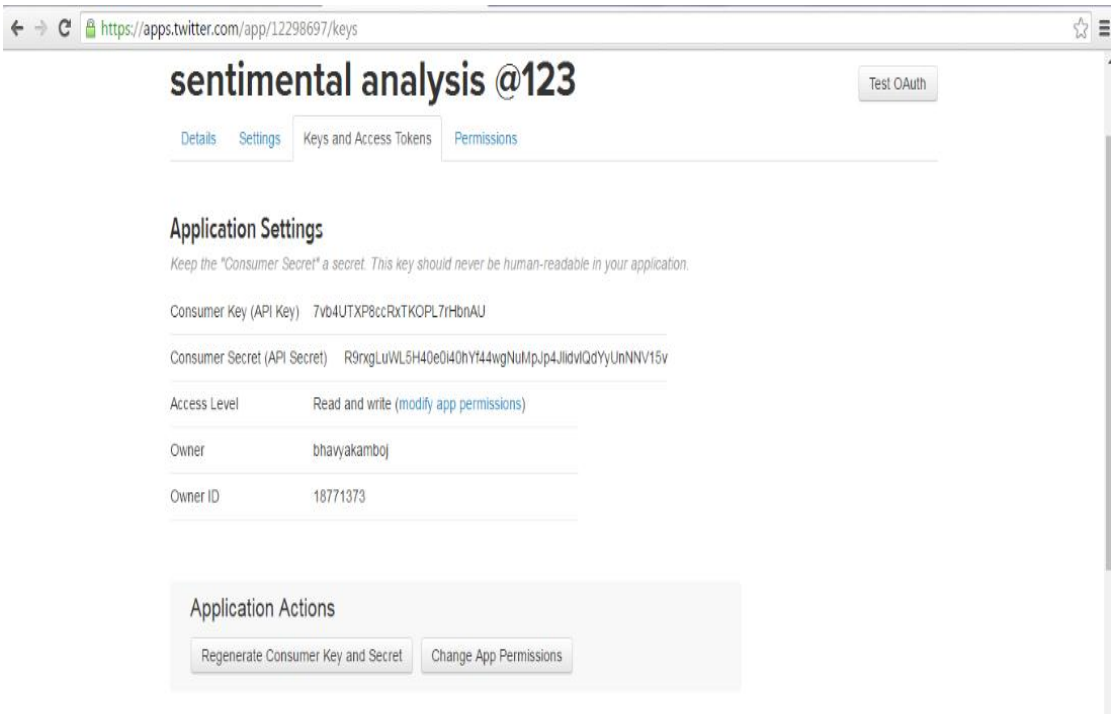


Fig 4.1 Data collection using Twitter API

2. **Data Cleaning** – All kinds of unwanted and repetitive tweets are removed and a training set is prepared. This training set consists of tweets segregated into three label folders: positive, negative and neutral. The training set information is given in the following chapter:

	Total
#Total Tweets	80000
#Repetitive	20000
#Total processed words	13,435,231 words

Table 1. About Training Set

The following steps are carried out in the data cleaning phase:

- **Text Data Cleaning-** All kinds of unwanted tweets were removed from the text.

- **No Repetition-** Repetition of a word may give emphasis on a particular feeling, so it is removed.
- **Text Correction-** A spell checker checks for misspelled words.

Figure 4.1 shows the basic methodology used in sentiment analysis of tweets based on airtel.

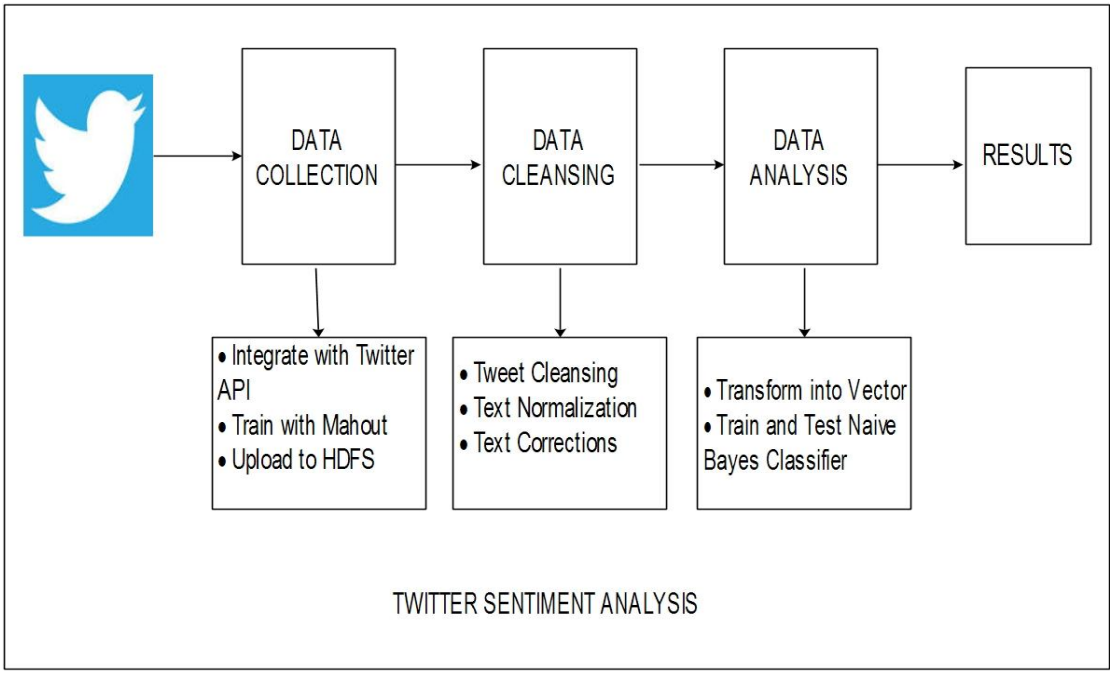


Fig 4.2 Sentiment analysis methodology in Mahout

The Mahout commands and results used for sentiment analysis have been discussed in Chapter 5.

4.3 METHODOLOGY II: CLUSTERING TWEETS USING FUZZY CMEANS CLUSTERING

The training set of tweets is subjected to fuzzy c means clustering by first converting the tweets to a sequence file which is then processed to form sparse vectors. Based on the tfidf weights i.e the term frequency inverse document frequency, fuzzy c means clustering is applied to tweets using CosineDistanceMeasure of Mahout. The algorithm for the proposed system is given below:

Algorithm:

Step: 1 Collect data related to airtel from twitter etc.

Step: 2 Remove the data which is not required for review analysis such as reviewer's name review date etc.

Step: 3 Identify some synonyms of different words and design the matrix.

Step: 4 Convert the tweets to a sequence file so that it forms a <key,value> pair.

Step: 5 To find the importance of the term by using TF-IDF method. Weight can be assigned of each term i as

$$Wi = tfi * \log (D/dfi)$$

tfi is the term frequency of the term

D is the number of documents

dfi is the number of documents in which term is present.

Sparse vectors are created from sequence files.

Step: 6 On the basis of their weight and score fuzzy c-means clustering is applied.

a) Random selection of k point as the initial centroids.
b) Find the distance of each point from each of the the centroid.
c) Create the membership matrix
d) Total membership for a point in all the clusters must add to 1.
e) Generate new centroid for each cluster with iteration all these steps.

Iteration will stop if centroids will be same as previous.

Step: 7 New clusters will be created after applying fuzzy clustering. Hence the top terms and similarity values between them are obtained for sentiment analysis.

Chapter 5: Experiment Results

5.1 Tweet Collection using Python script

Twitter is a widely used social platform used to post comments on various topics in the form of short status messages. In this thesis, tweets have been collected and converted into a training set by using a python script. The tweets are collected by using hashtag @airtel_presence, which are meant to judge the satisfaction and dissatisfaction level of customers with respect to the service provider.

negative	705807952	#ASCI upholds complaints for misleading adshttps//t.co/Z4I8VaaX84
negative	705807933	RT @Khem_Raj #AirtelSucks #Airtel #Sucks @Airtel_Presence @airtelindia
neutral	705807933	RT @priyamani6 And so my cameo appearance on #Airtel super singer with my aunt is airing today!930pm..@vijaytelevision https
negative	705791653	I wish to open a page on #airtel network.. pathetic service ever witnessed @airtelindia
negative	705785891	@Airtel_Presence shared my contact. Hope some quick resolution is provided. Tired of #Airtel #poorservice Need action asap
neative	705775222	RT @Khem_Raj #AirtelSucks #Airtel #Sucks @Airtel_Presence @airtelindia
negative	705768131	@VodafoneIN @Vodafone_CEO @rsprasad @GoI_DeitY Sir, is this how they are allowed to operate? Like Petty Thiefs! Need #idea #Airtel #BSNL
neutral	705764752	Make OLA's [@onabajoolawale] Hit Songs Your Caller RingBack Tunes https //t.co/RnOPn52yw2 on #MTN #Airtel... https
neutral	705763622	Make OLA's [@onabajoolawale] Hit Songs Your Caller RingBack Tunes https //t.co/3OEg8dhTuk on #MTN #Airtel #Etisalat https
neutral	705763612	Make OLA's [@onabajoolawale] Hit Songs Your Caller RingBack Tunes https //t.co/RnOPn52yw2 on #MTN #Airtel #Etisalat https
positive	705754692	8 résidus alimentaires excellents pour ta santé ! Tout sur https//t.co/12siqjj0G0 #airtel #rdc
positive	705753421	Ces résidus alimentaires sont excellents pour ta santé. Retrouve-les ici https//t.co/DZgrIzEB5w #airtel #tchad
positive	705752111	Nutrition déchets alimentaires parfaits pour ta santé. La liste ici https//t.co/FwWQO6LPaU #airtel #burkina
positive	705750912	Sais-tu que certains résidus alimentaires sont excellents pour ta santé ? Découvre-les ici https
positive	705749643	Voici 8 résidus alimentaires excellents pour ta santé ! A découvrir ici https//t.co/S5N8lqMq8t #airtel #congo
positive	705748383	Nutrition 8 résidus alimentaires excellents pour ta santé ! A découvrir ici https //t.co/QtUGdrD7GB #airtel #niger
positive	705747133	Voici 8 résidus alimentaires excellents pour ta santé ! A découvrir ici https//t.co/XNFmlpRo5L #airtel #madagasikara
negative	705807953	#ASCI upholds complaints for misleading adshttps//t.co/Z4I8VaaX84
negative	705801932	RT @Khem_Raj #AirtelSucks #Airtel #Sucks @Airtel_Presence @airtelindia
neutral	705887933	RT @priyamani6 And so my cameo appearance on #Airtel super singer with my aunt is airing today!930pm..@vijaytelevision https
negative	705791654	I wish to open a page on #airtel network.. pathetic service ever witnessed @airtelindia
negative	705785894	@Airtel_Presence shared my contact. Hope some quick resolution is provided. Tired of #Airtel #poorservice Need action asap
neative	705775233	RT @Khem_Raj #AirtelSucks #Airtel #Sucks @Airtel_Presence @airtelindia
negative	705768133	@VodafoneIN @Vodafone_CEO @rsprasad @GoI_DeitY Sir, is this how they are allowed to operate? Like Petty Thiefs! Need #idea #Airtel #BSNL
neutral	705764734	Make OLA's [@onabajoolawale] Hit Songs Your Caller RingBack Tunes https //t.co/RnOPn52yw2 on #MTN #Airtel... https
neutral	705763613	Make OLA's [@onabajoolawale] Hit Songs Your Caller RingBack Tunes https //t.co/3OEg8dhTuk on #MTN #Airtel #Etisalat https
neutral	705763612	Make OLA's [@onabajoolawale] Hit Songs Your Caller RingBack Tunes https //t.co/RnOPn52yw2 on #MTN #Airtel #Etisalat https
positive	705754244	8 résidus alimentaires excellents pour ta santé ! Tout sur https//t.co/12siqjj0G0 #airtel #rdc
positive	705753424	Ces résidus alimentaires sont excellents pour ta santé. Retrouve-les ici https//t.co/DZgrIzEB5w #airtel #tchad
positive	705752165	Nutrition 8 déchets alimentaires parfaits pour ta santé. La liste ici https//t.co/FwWQO6LPaU #airtel #burkina
positive	705750903	Sais-tu que certains résidus alimentaires sont excellents pour ta santé ? Découvre-les ici https
positive	705749633	Voici 8 résidus alimentaires excellents pour ta santé ! A découvrir ici https//t.co/S5N8lqMq8t #airtel #congo
positive	705748384	Nutrition 8 résidus alimentaires excellents pour ta santé ! A découvrir ici https //t.co/QtUGdrD7GB #airtel #niger

Fig 5.1 .Tweets collected using Python script.

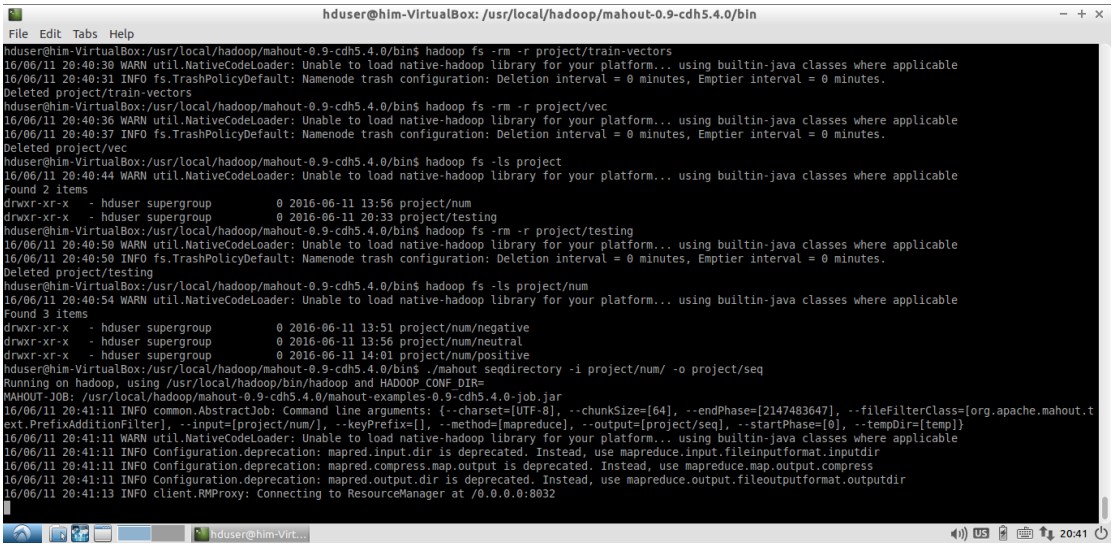


Fig 5.2 Uploading training set on HDFS (in the project folder)

5.2 EXPERIMENT: DATA ANALYSIS

After the training set has been prepared, data is analyzed by uploading it on HDFS and Naïve Bayes classification is carried out using Mahout commands.

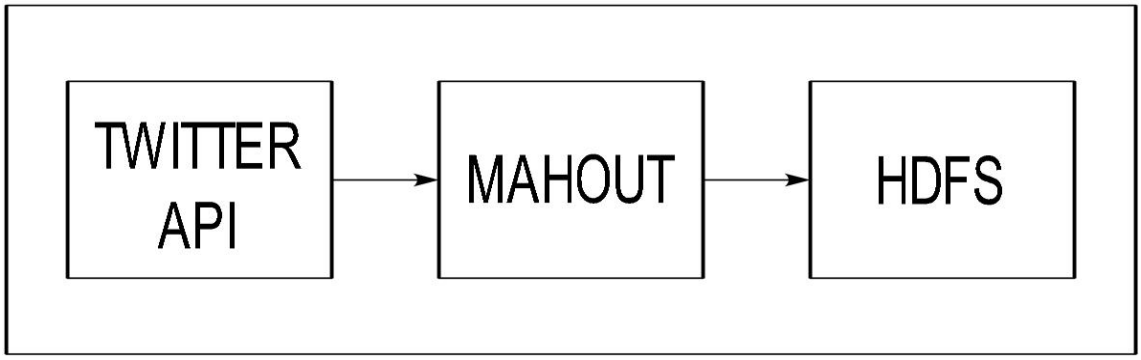


Fig 5.3 Uploading tweets to HDFS

The commands used are shown in the following steps.

i. **Create a new directory on hdfs named tweets :**

hadoop fs -mkdir tweets

ii. **Load training set on hdfs**

hadoop fs -put home/trainingset-folder tweets

iii. **Create into < Text, Text > sequence file:**

$home/usr/local/hadoop/mahout-0.9-cdh5.4.0/bin> $ ./mahout seqdirectory -i tweets -o tweets/tweetseq –ow

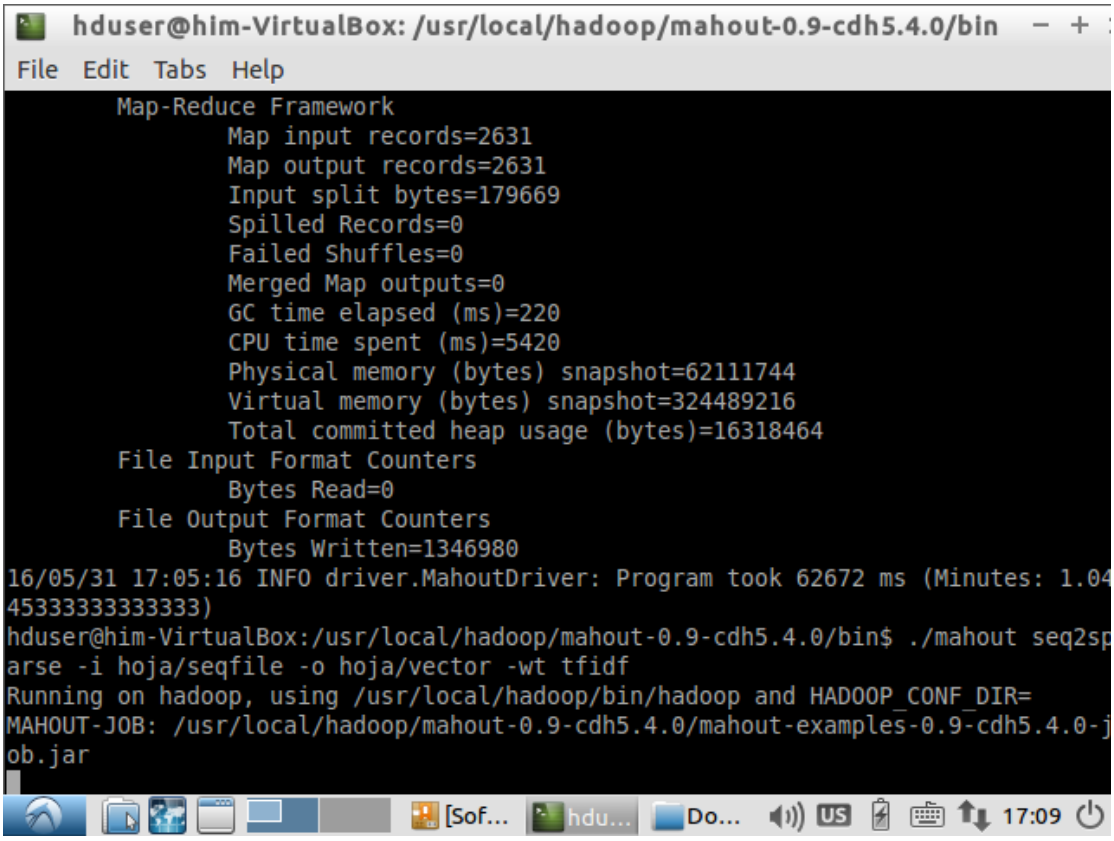


Fig 5.4 Converting training set to sequence file

iv. Convert and preprocesses the dataset into a < Text, VectorWritable > SequenceFile containing term frequencies for each document.

$home/usr/local/hadoop/mahout-0.9-cdh5.4.0/bin> $ ./mahout seq2sparse -i tweets/tweetseq -o tweets/tweetvectors -wt tfidf

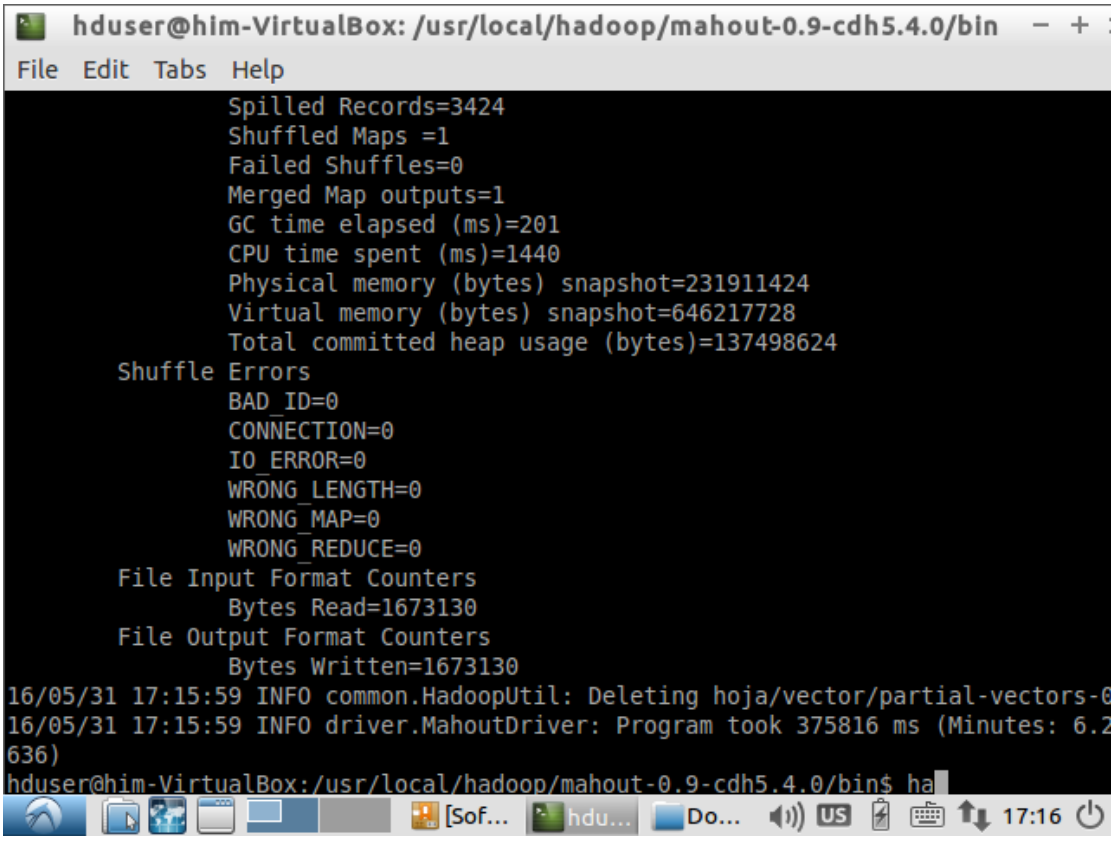


Fig 5.5 Converting sequence files to vector files

The training set is thus converted into vectors using *tfidf weight (*term frequency x document frequency). This is done to assign weight to every term in the word list created from the set. This could be done by calculating the Term frequency, which finds the frequency of each term in the document. The second aspect is Inverse Document Frequency which means that the lesser the presence of a term in all documents, the more is the term value or weight in this matter . The following files are generated in HDFS:

- **df-count** : sequence file with mapping of word id to number of documents.
- **dictionary.file-0**:sequence file mapping word to word id
- **frequency.file-0**:mapping word id to word count
- **tf vectors**:sequence file having frequency for each document
- **tfidf-vectors**:sequence file mapping document id to tfidf weight for each word in the document
- **tokenized-documents**:sequence file with association of document id and list of words
- **wordcount**:sequence file with word to word count association.

v. **Split the preprocessed dataset into training and testing sets.(Training set is 70% of the collected tweets)**

$home/usr/local/hadoop/mahout-0.9-cdh5.4.0/bin> $ ./mahout split -i tweetvectors/tfidf-vectors --trainingOutput tweets/train-vectors --testOutput tweets/test-vectors --randomSelectionPct 30 --overwrite --sequenceFiles -xm sequential

vi. **Train the classifier.**

$home/usr/local/hadoop/mahout-0.9-cdh5.4.0/bin> $./mahout trainnb -i tweets/train-vector -el

The training set trains the classifier. It creates a model comprising of the matrix with word id and the corresponding label id, and a label index which consists of an association label and its label id. Mahout commands are used to test the classifier on the training and testing set.

vii. **Test the classifier**

$home/usr/local/hadoop/mahout-0.9-cdh5.4.0/bin> $ ./mahout testnb -i tweets/test- vectors -m tweets/model -l tweets/labelindex -ow -o tweets/testing –c

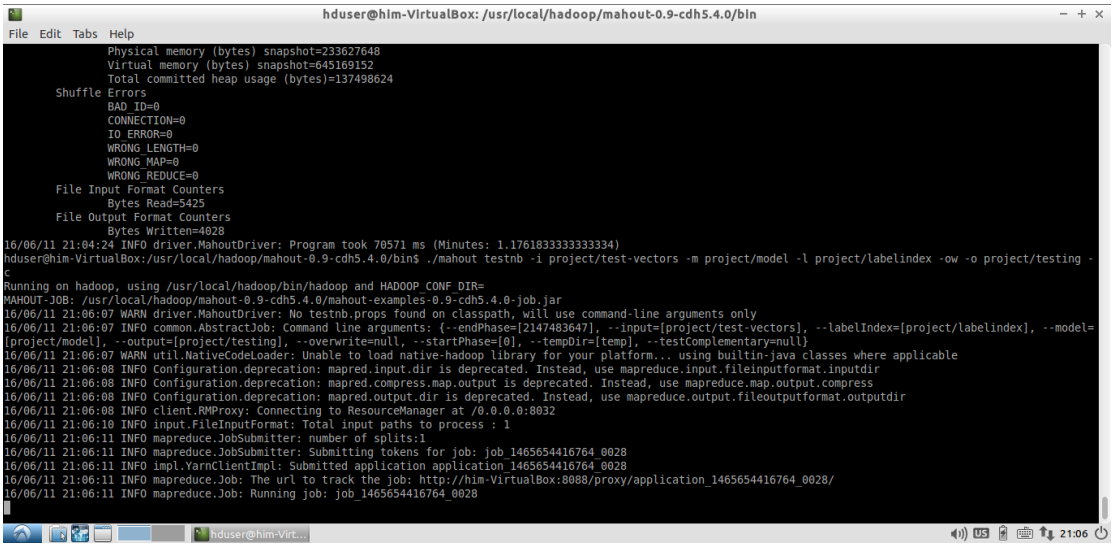


Fig 5.6 Testing the classifier

5.3 RESULTS OF SENTIMENT ANALYSIS

In the classification step, 70% of the dataset which comprises of 80,000 tweets, is randomly selected for training set. Testing is thus carried on the remaining 30% of the dataset.

The analysis showed that around 20% of the opinions were negative, 38% were neutral and 40% were positive for the given two months, for the training set. Fig 5.4 shows the results obtained after Naïve Bayes classification.

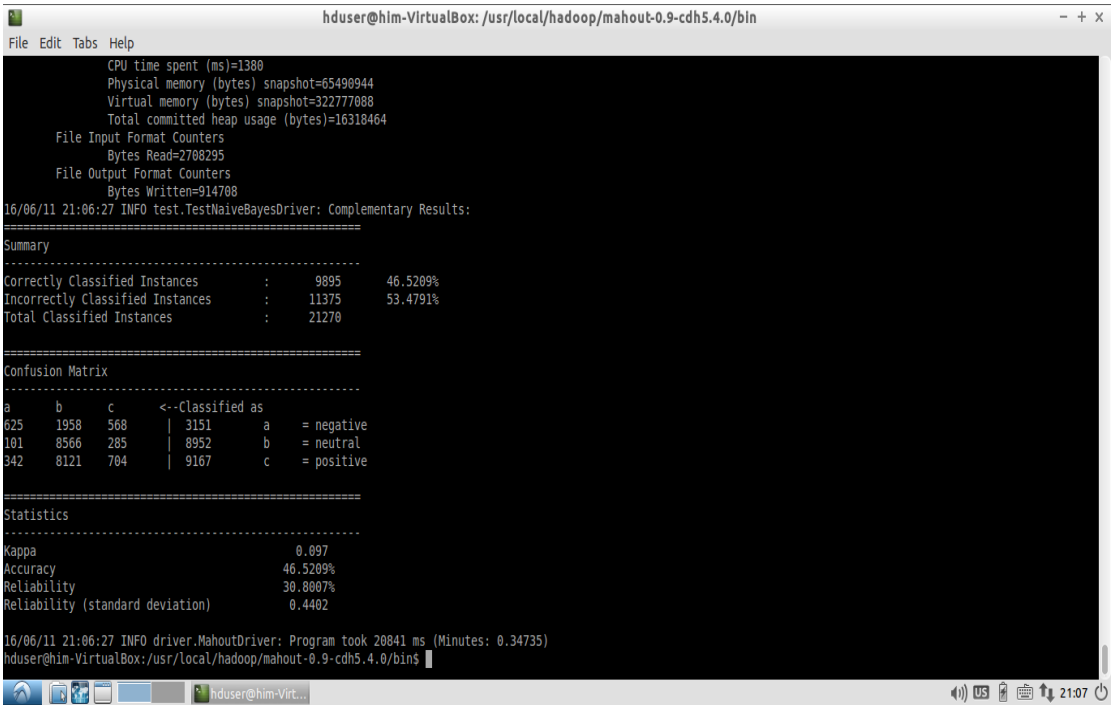


Fig 5.7 Naïve Bayes classification on testing set of tweets

Fig 5.7 shows a column graph of the increasing and decreasing trend in the judgments of users, as inferred from the results obtained by carrying out the classification. Thus more users hold positive reviews for airtel i.e. about 40%. Around 38% views are neutral and negative views are the least for this service provider.

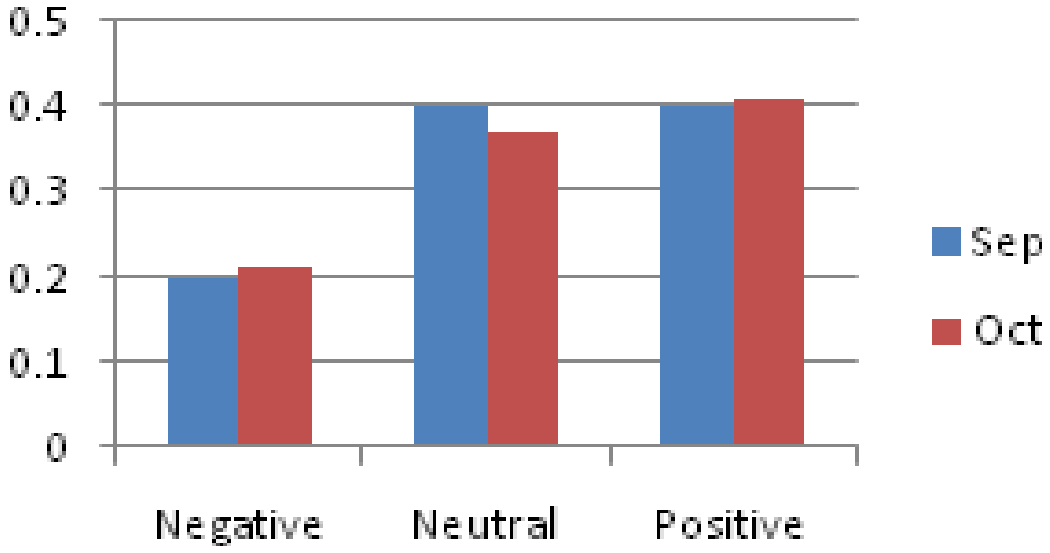


Fig 5.8 Sentiment analysis of users based on classification results.

The statistics of results are shown in the graph below:

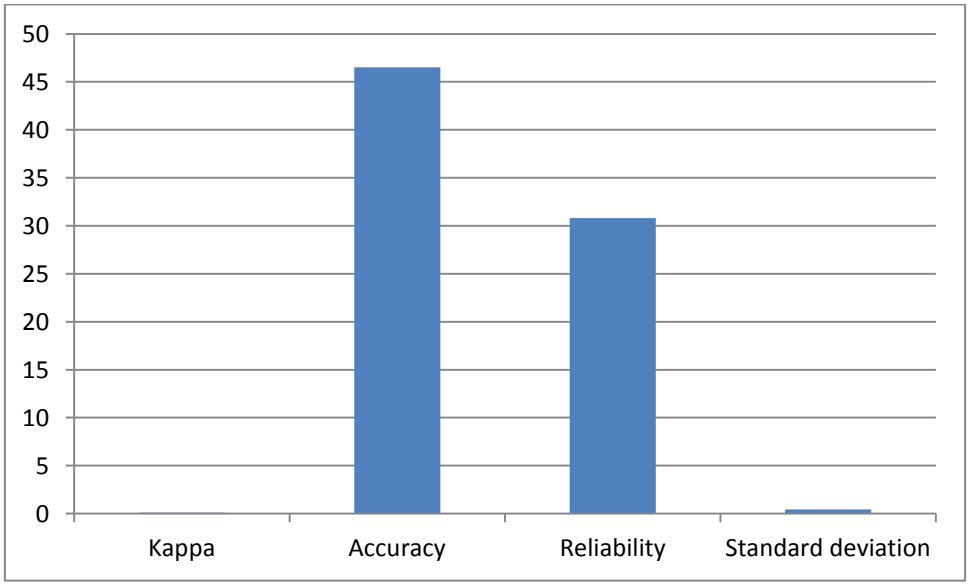


Fig 5.9. Statistics of Naïve Bayes classification

5.4 FUZZY CLUSTERING OF TWEETS

The training set of tweets is clustered by using fuzzy c means clustering.This is done by using the following steps:

i. Convert the input text to sequences:

The first step is to convert the input documents into a sequence form using seqdirectory . This is a necessary step before being able to perform fuzzy cmeans clustering. The tweets are stored in a directory named tweets uploaded toHDFS.

$mahout seqdirectory -i monu/tweets -o monu/sequencefile-c UTF-8 -chunk

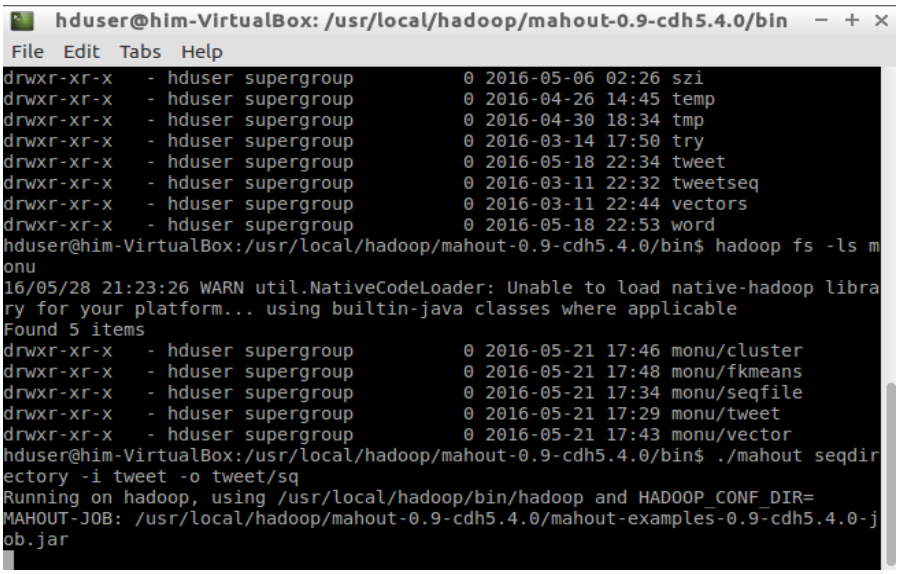


Fig 5.10. Convert tweets to sequence file

ii. Convert the generated sequences to sparse vectors using seq2sparse

It is necessary to obtain vectors from sequence file to use it as part of the input to the clustering process.

$mahout seq2sparse -i monu/sequencefile -o monu/vectorfile --maxDFPercent 85 –namedVector

iii. Run fuzzy c-means clustering on tweets

We now run the fuzzy c-means clustering by giving initial number of clusters and membership values greater than 1.

$mahout fkmeans -i ~/monu/vectorfile/tfidf-vectors/ -c monu/clusterfile -o monu/fuzzy -dm

org.apache.mahout.common.distance.CosineDistanceMeasure -x 10 –m 1.5 -k 20 -ow --clustering –cl

iv. Dump the results to file

In order to view the clustering results, it needs to be dumped through the use of a special program called clusterdump.

$mahout clusterdump –i monu/fuzzy/clusters-*-final –d monu/vectorfile/dictionary.file-0 -dt sequencefile

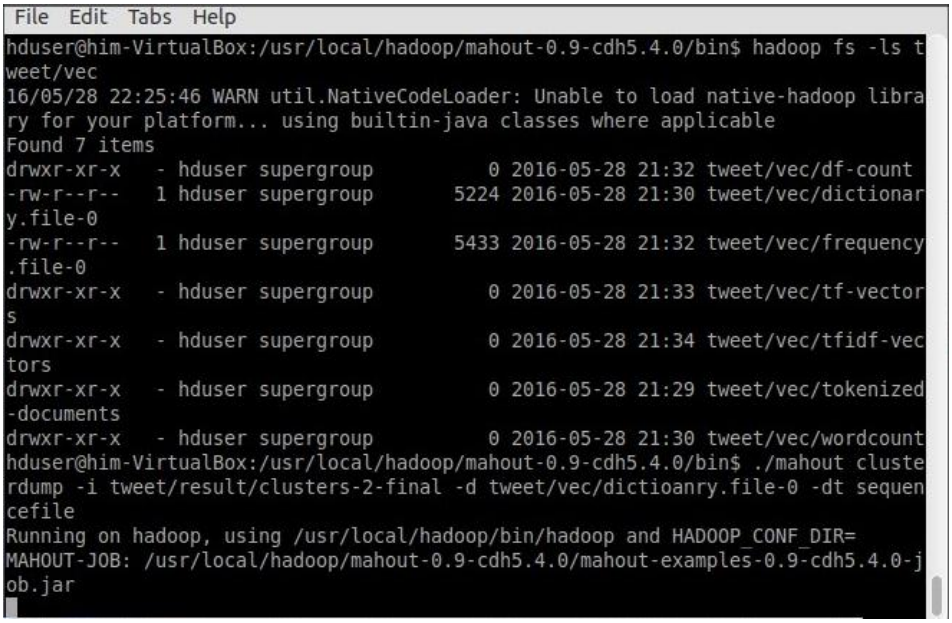


Fig 5.11. Dumping cluster results to a file

5.5 SENTIMENT ANALYSIS USING CLUSTERING

The results obtained are evaluated on the basis of the similarity between words to create a lexicon. Based on this lexicon, sentiment values are obtained for every emotion. Thus fuzzy clustering helps in classification of emotions in an effective way. Fuzzy clustering is done on the testing set of tweets used in Naïve Bayes classification.

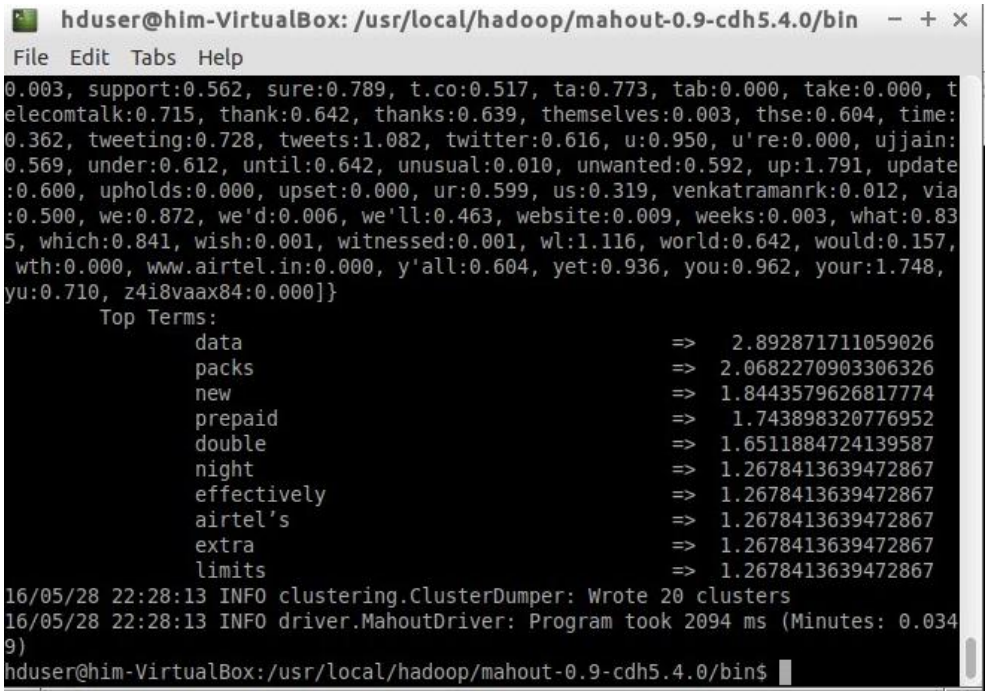


Fig 5.12. Clusters showing top terms of tweets

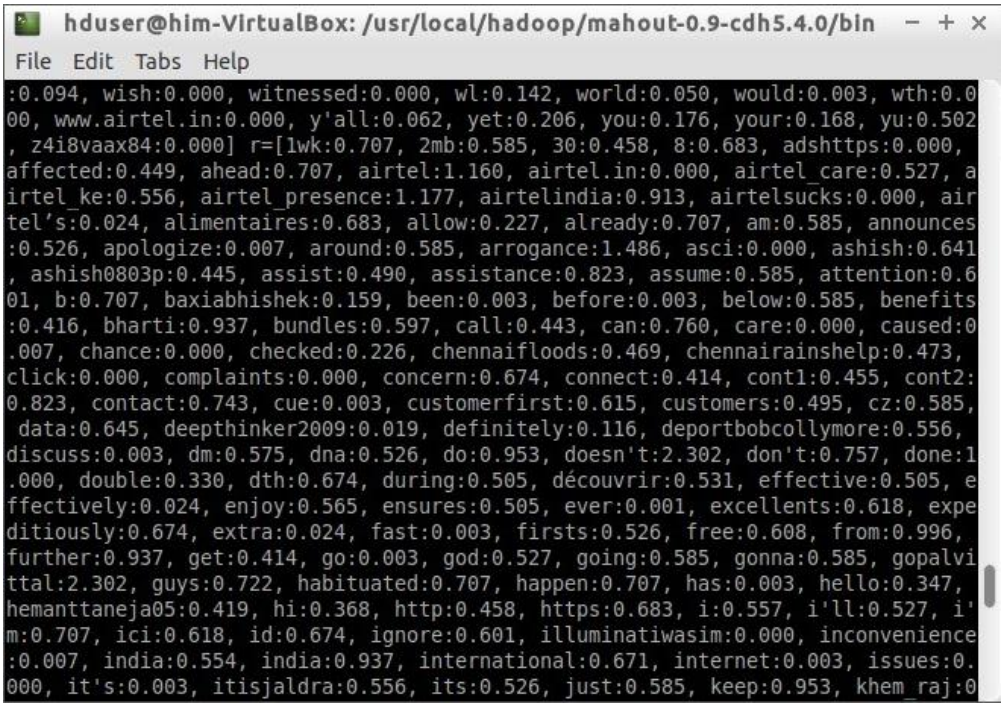


Fig 5.13. Similarity between words helps to analyse sentiments

Thus with the help of classification and clustering, it is possible to analyze sentiments of people, in our case for service provider airtel.

5.6 COMPARISON BETWEEN NAÏVE BAYES CLASSIFICATION AND CLUSTERING

It is observed that Naïve Bayes classification takes lesser time as compared to fuzzy c-means clustering to analyze tweets. This is shown in the graph below:

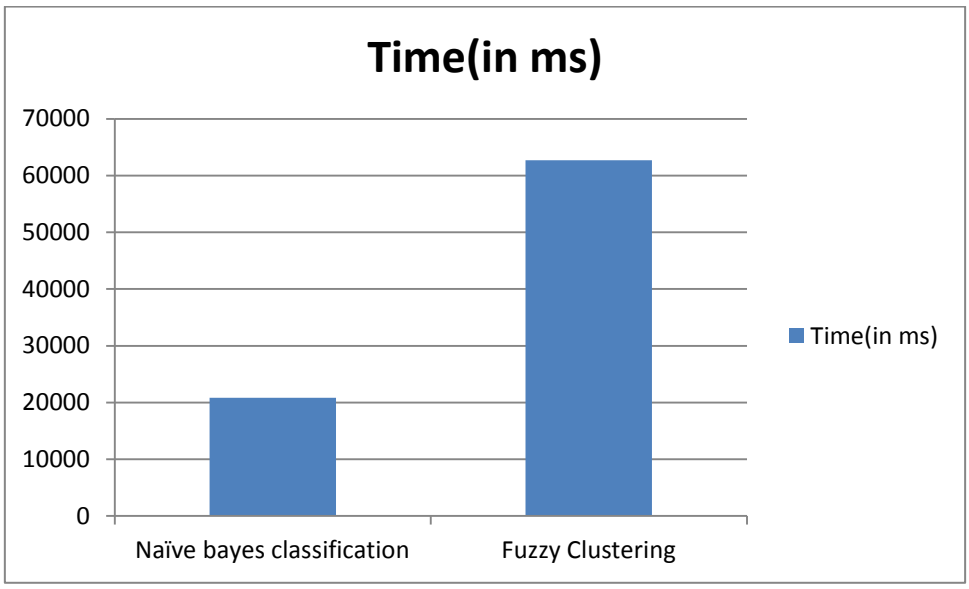


Fig 5.14. Time taken to classify and cluster tweets

Chapter 6: Conclusion and Future Scope

In the thesis, various opinions have been analysed based on tweets of Airtel data. It is also useful in gauging what people feel when it comes to diverse topics related to any field. The number of users who have positive, negative and neutral reviews about Airtel help to understand the pattern of satisfaction related to this service provider in the nation. Since data was huge, it was trained and tested using a Big data platform. It would have been difficult to analyze so many opinions on the basis of conventional data analytics tools. Further, the data is also clustered using fuzzy c-means clustering to analyse the sentiments of users with the help of top terms and similarity between them.

As future work, we can further compare the services of various providers and judge which one is the best. Using Hash tags, we can provide a simple automated method to evaluate what people think. Thus collecting information from social networks and analyzing it using Big Data techniques has left behind the traditional database approach.

References

[1] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up Sentiment Classification using Machine Learning Techniques". In Proceedings of the Empirical Methods on Natural Language Processing, Pennsylvania, 2002, pp. 79-86.

[2] G. Vinodhini and R. Chandrasekaran, “Sentiment analysis and opinion mining: A survey,” International Journal, vol. 2, no. 6, 2012.

[3] Y. Mejova, “Sentiment analysis: An overview,” Comprehensive exampaper,http://www.csuioedu/˜ymejova/publications/CompsYelenaMejova. pdf [2010-03], 2009.

[4] [Rui Xia, Chengqing Zong, Shoushan Li, “Ensemble of feature sets and classification algorithms for sentiment classification”, Information Sciences 181 (2011) 1138–1152.

[5] H. Cui, V. Mittal, and M. Datar, “Comparative Experiments on Sentiment Classification for Online Product Reviews.” In Proceedings of AAAI-06, 2006, pp.1265-1270.

[6] P. Domingos and M. Pazzani, “On the optimality of the simple Bayesian classifier under zero-one loss,” Machine Learning, vol. 29, no. 2-3, pp. 103–130, 1997

[7] Z. Niu, Z. Yin, and X. Kong, “Sentiment classification for microblog by machine learning,” in Computational and Information Sciences (ICCIS), 2012 Fourth International Conference on, pp. 286–289, IEEE, 2012.

[8] Pak and P. Paroubek, “Twitter as a corpus for sentiment analysis and opinion mining,” in Proceedings of LREC, vol. 2010, 2010.

[9] R. Xia, C. Zong, and S. Li, “Ensemble of feature sets and classification algorithms for sentiment classification,” Information Sciences: an International Journal, vol. 181, no. 6, pp. 1138–1152, 2011.

[10] Neethu M S and Rajasree R, “Sentiment Analysis in Twitter using Machine Learning Techniques” 4th ICCCNT 2013 July 4 - 6, 2013, Tiruchengode, India IEEE – 31661.

[11] J. Kamps, M. Marx, R. J. Mokken, and M. De Rijke, “Using wordnet to measure semantic orientations of adjectives,” 2004.

[12] C. Fellbaum, “Wordnet: An electronic lexical database(language,speech, and communication),” 1998.

[13] P. D. Turney, “Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews,” in Proceedings of the 40th annual meeting on association for computational linguistics, pp. 417–424, Association for Computational Linguistics, 2002.

[14] Y. Xia et al, “Word polarity disambiguation using Bayesian model and opinion level features” Cognitive Computation, vol. 7,no.3,2015.

[15] M.Dragoni, A.G. Tettamanzi and C.da Costa Pereira, “Combined system for concept-level sentiment analysis” Semantic web evaluation challenge, springer,2014,pp,21-27.

[16] M.Araujo “iFeel: A System that compares and combines sentiment analysis methods,” Proc.23 International Conf. World Wide Web,2014,pp.75-78.

[17] J.M Chenlo and D.E.Losada, “An Empirical Study of Sentence Features for Subjectivity and Polarity Classification”, Information Sciences, vol.280,2014,pp.275-288.

[18] J.K, C.Chung ,C.E.Wu, and R.T.H.Tsai, “Improve Polarity Detection of Online Reviews with bag of Sentiment Concepts”Proc,11 European Semantic Web Conferrence,2014.

[19] F.Bravo-Marquez, M.Mendoza, and B.Poblete, “Meta-Level Sentiment Models for Big Social Data Analysis,” Knowledge-Based Systems, vol.69,2014,pp,86-99.

[20] G,Gezici , “SU-Sentilab: A Classification System for Sentiment Analysis in Twitter,” Proc. International Workshop Semantic Evaluation,2013, pp.471-477.

[21] D.R.Recupero, “Sentilo: Frame-Based Sentiment Analysis,” Cognitive Computation, vol.7,no.2,2014,pp.211-225.

[22] Hussam Hamdan, Frederic Béchet and Patrice Bellot Experiments with DBpedia, WordNet and SentiWordNet as resources for sentiment analysis in micro-blogging Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013), pages 455–459, Atlanta, Georgia, June 14-15, 2013. c 2013 Association for Computational Linguistics

[23] Erik cambria, ieee computer society 1541-1672/16 @2016.

[24] S. Lloyd, least squares quantization in PCM, ieee transactions On information theory 28(2): 129-137

[25] X. Su, G. Gao, Y. Tian, "A Framework to Answer Questions of Opinion Type" Seventh Web Information Systems and Applications Conference in 2010.

[26] Timothy L. Acorn, Sherry H. Walden "Smart: Support: Management Automated Reasoning Technology for Compaq Customer Service" In proceeding of: Proceedings of the Fourth Conference on Innovative Applications of Artificial Intelligence.

[27] Recio1,J A. Agudo1,B D. G´omez-Martin1,M A and Wiratunga2,N., "Extending COLIBRI for Textual CBR", Proceedings of the Sixth Spanish Conference on Programming and Languages, (SCPL'05)

[28] E. Rashid, S. Patnayak and V. Bhattacherjee, "A Survey in the Area of Machine Learning and Its Application for Software Quality Prediction" ACM SIGSOFT Software Engineering Notes Volume 37 Issue 5, September 2012.

[29] Gang Li, Fei Liu, "A Clustering-Based Approach On Sentiment Analysis", 978-1-4244-6793-8/10 ,IEEE,2010.

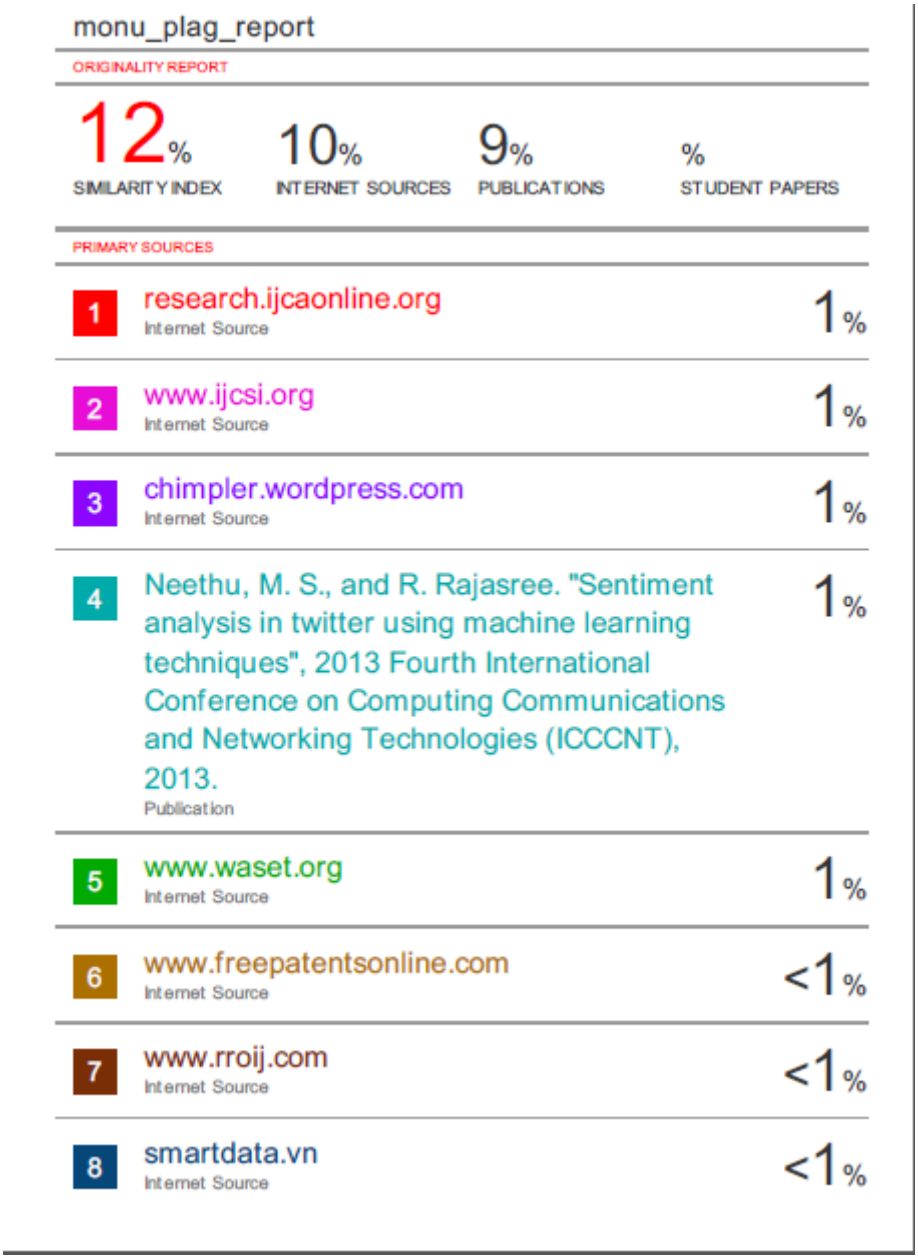
List of Publications

Monu Kumar, Dr. Anju Bala, "'**ANALYZING TWITTER SENTIMENTS THROUGH BIG DATA'** has been accepted in 'Proceedings of the 10th INDIACom:3rd 2016 International Conference on Computing For Sustainable Global Development' which is scheduled from 16th -18th March, 2016 at Bharati Vidyapeeth's Institute Of Computer Applications and Management.

Video URL

https://www.youtube.com/channel/UCPzbkGjRO_Co9lz8nQwkAIA?guided_help_flow=3

Plagiarism Report



monu_plag_report

ORIGINALITY REPORT

12%	10%	9%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	research.ijcaonline.org Internet Source	1%
2	www.ijcsi.org Internet Source	1%
3	chimpler.wordpress.com Internet Source	1%
4	Neethu, M. S., and R. Rajasree. "Sentiment analysis in twitter using machine learning techniques", 2013 Fourth International Conference on Computing Communications and Networking Technologies (ICCCNT), 2013. Publication	1%
5	www.waset.org Internet Source	1%
6	www.freepatentsonline.com Internet Source	<1%
7	www.rroij.com Internet Source	<1%
8	smartdata.vn Internet Source	<1%