SENTIMENT ANALYSIS FOR SOCIAL MEDIA AND ONLINE REVIEW

A dissertation submitted

By

Rajni Singh

To

Department of Computer Science and Engineering

In partial fulfillment of requirement for the

Award of the degree of

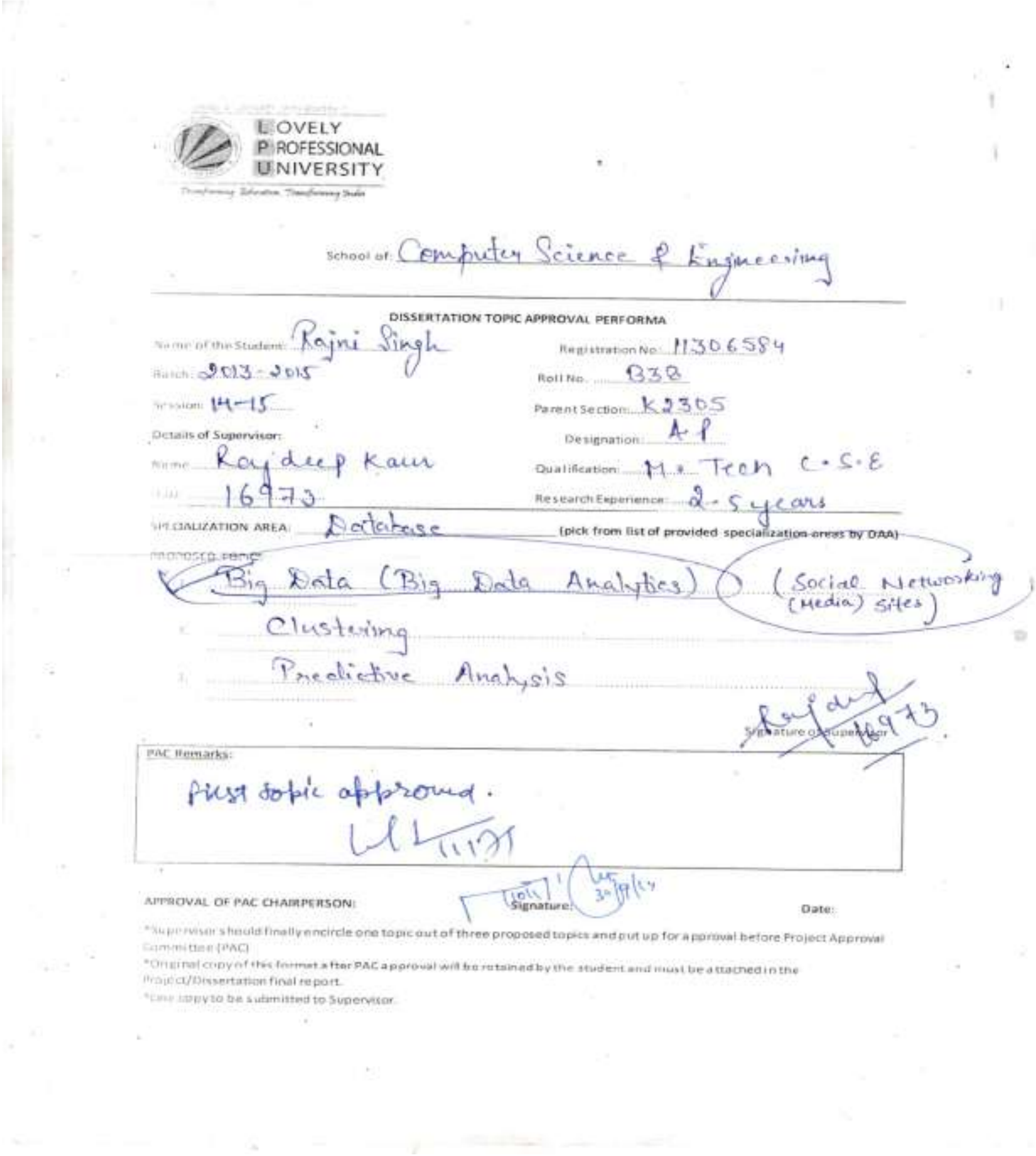
Master of Technology in Computer Science

Under the guidance of

Rajdeep Kaur

(May 2015)

PAC APPROVAL



LOVELY
PROFESSIONAL
UNIVERSITY

School of: Computer Science & Engineering

DISSERTATION TOPIC APPROVAL PERFORMA

Name of the Student: Rajni Singh	Registration No: 11306584
Batch: 2013-2015	Roll No. B38
Session: 14-15	Parent Section: K2305
Details of Supervisor:	Designation: A P
Name: Rajdeep Kaur	Qualification: M. Tech C.S.E
UID: 16973	Research Experience: 2.5 years

SPECIALIZATION AREA: Database (pick from list of provided specialization areas by DAA)

PROPOSED TOPICS

1. Big Data (Big Data Analytics) (Social Networking (Media) sites)
2. Clustering
3. Predictive Analysis

Rajdeep 16973
Signature of Supervisor

PAC Remarks:

first topic approved.

APPROVAL OF PAC CHAIRPERSON: Signature: 30/9/14 Date:

*Supervisor should finally encircle one topic out of three proposed topics and put up for approval before Project Approval Committee (PAC)

*Original copy of this format after PAC approval will be retained by the student and must be attached in the Project/Dissertation final report.

*One copy to be submitted to Supervisor.

ABSTRACT

Social media monitoring has been growing day by day so analyzing of social data plays an important role in knowing customer behavior. So we are analyzing Social data such as Twitter Tweets using sentiment analysis which checks the attitude of User review on movies. So our aim is to develop a dictionary based on social media keywords and find hidden relationship pattern from these keyword. This provide hidden relation between different keywords and a dictionary of the keywords on the basis of categories of different comments & tweets.

CERTIFICATE

This is to certify that Rajni Singh has completed M.Tech dissertation titled “Sentiment for social media and online review” under my guidance and supervision. To the best of my knowledge the present work is the result of her original investigation and study. No part of the dissertation proposal has ever been submitted for any other degree or diploma.

The dissertation proposal is fit for the submission and the partial fulfillment of the conditions for the award of M.Tech computer science and Engineering.

Date: ______________

Signature of supervisor:

Name:

UID:

ACKNOWLEDGEMENT

First and foremost I would like to thank almighty for giving me courage to bring up this Dissertation. Before getting into thick and thin of this Dissertation I would like to show my gratitude to some of the people who have helped me in this project. Firstly I would like to purpose a word thanks to my mentor RAJDEEP KAUR who has encouraged me to get through this Dissertation. Secondly I would like to thanks my friends who gave me unending support and helped me in numerous ways from the stage when the idea of the thesis was conceived. I am very thankful to all of them for making my work complete successfully under their guidance.

DECLARATION

I hereby declare that the dissertation entitled, "**Sentiment analysis for social media and online review"** submitted for the M.Tech. Degree is entirely my original work and all ideas and references have been duly acknowledged. It does not contain any work for the award of any other degree or diploma.

Date: 30th April 2015

Rajni Singh

RegNo.:11306584

Table of Contents

Topic **Page No.**

LIST OF TABLE

Table name **Page no.**

LIST OF FIGURES

Figure name **Page no.**

CHAPTER 1

INTRODUCTION

Nowadays, Social media is becoming more and more popular since mobile devices can access social network easily from anywhere. Therefore, Social media is becoming an important topic for research in many fields. As number of people using social network are growing day by day, to communicate with their peers so that they can share their personal feeling everyday and views are created on large scale.

Social Media Monitoring or tracking is most important topic in today's current scenario. In today many companies have been using Social Media Marketing to advertise their products or brands, so it becomes essential for them that they can be able to calculate the success and usefulness of each product [2].

For Constructing a Social Media Monitoring, various tool has been required which involves two components: one to evaluate how many user of their brand are attracted due to their promotion and second to find out what people thinks about the particular brand.

Humors, that have been generated can be evaluated usually by performing various Key performance factors such as the number of followers or friends, the number of likes or shares or comment for each post and more difficult one like engagement rate, response time to evaluate them and other composite measures. Measuring the Large dataset is usually direct and can be done by using some statistical method.

On the other hand, to evaluate the opinion of the users is not as easy as it seems to all users. For evaluating their attitude may requires to perform Sentiment Analysis, which is defined as to identify the polarity of customer behavior, the subjective and the emotions of particular document or sentence. To process this we need Machine Learning and Natural Language Processing methods and this is place where most of the developers facing difficulty when they are trying to form their own tools.

Over the recent years, an emerging interest has been occurred in supporting social media analysis for advertising, opinion analysis and understanding community cohesion. Social media data adapts to many of the classifications attributed for “big-data” – i.e. volume, velocity and variety. Analysis of Social media needs to be undertaken over large volumes of data in an efficient and timely manner.

Analysing the media content has been centralized in social sciences, due to the key role that the social media plays in modelling public opinion. This type of analysis typically on the preliminary coding of the text being examined, a step that involves reading and annotating the text and that limits the sizes of the data that can be analysed.

Social media platforms are refers to the websites and applications that enable people for interacting, building, sharing and exchange information. Every day total world post 400 million tweets on Twitter, 350 million photos on Face book and 4 billion videos on YouTube. This has encouraged developer to develop of new techniques and methodological approaches for capturing, processing and analyzing large and complex data. Big data approaches for analyzing social media data can increase understanding of how people thoughts and act towards a particular topic. Companies can use this information and try to influence and knows users’ behaviors in the future [3].

1.1 Collecting Social Media Data

Accessing of large quantity of social data on millions of people’s as to track their activities and behaviors that are needed for researchers and organizations. Social media data which have not been formed for the purpose of research, so they can view insight into the people attitude that interact online. Huge amount of Data can be automatically extracted from social media websites via Application Programming Interfaces.

1.2 Examples of Social Media Platform

- Face book: a social networking service that allow the user to create their personal profile. For interaction it takes the form of posts and the user can point out their preferences for the user generated comments, content, articles, product and services.
- YouTube: allow the user to upload, comment and view on the videos
- Twitter: a micro blogging services that allow the user to broadcast and read tweets of up to 120 characters.
- Flickr: allow the user to upload, comment and view on photos.

Huge amount of data can be accessed for research or a profit-making purpose that mainly depends on whether users have selected their content as public or private.

- **Public data.** It includes Twitter content. Everyone has the access for the existing data from various social media platforms.
- **Private data.** Here Facebook content or data is considered as private. Social media companies may transfer huge content that have been consider or thought to be private to third parties if users have given their identification, or if the data are anonymised.

1.3 Sentiment Analysis

Sentiment analysis refers to the use of natural language processing to identify and extract one-sided information in source materials or simply it refers to the process of detecting the polarity of the text. It also referred as opinion mining, as it derives the opinion, or the attitude of a user. A common approach of using this is described how people think about a particular topic.

Sentiment analysis helps in determining the thoughts of a speaker or a writer with respect to some subject matter or the overall contextual polarity of a document. The attitude may be his or her decision or estimate, the emotional state of the user while writing.

Sentiment Analysis can be used to determine sentiment on a variety of level. It will score the entire document as positive or negative, and it will also score the reaction of individual words or phrases in the document.

Sentiment Analysis can track a particular topic, many companies use it to track or observe their products, services or status in general. For example, if someone is attacking your brand on social media, sentiment analysis will score the post as enormously negative, and you can create alerts for posts with hyper-negative sentiment scores.

Sentiment Analysis is hard

Today, Sentiment analysis plays an important role where various machine learning technique is used in determining the sentiment of very huge amounts of text or speech. Various application tasks include such as determining how someone is excited for an upcoming movie, correlates different views for a political party with people's positive attitude towards vote for that party, or by converting written hotel reviews into 5-star based on scaling across categories like 'quality of food', 'services', 'living room' and 'facilities' provided.

As there is huge amount of information is shared on social media, forums, blogs, newspaper etc. it is easy to see why there is a need for sentiment analysis as there is much information to process manually which is not possible in today's time.

Main problem is that machine learning processes is not that much accurate. For a simple process to separate 'positive' view from 'negative' view on social media, many solutions can provide only performance around 80% accuracy. It can be useful to the track broad trends over time, but it may limit analysis of fined grained. Four main factors not to depend blindly on any tool for sentiment analysis:

- **Context**: A word can have positive or negative sentiment that can have the opposite implication that depends on context for e.g. "my service provider has done a *great* job when it comes to pinching money from me".

- **Sentiment Ambiguity**: a sentence which has a positive or negative word doesn't essentially express any sentiment. So, Sentences without any sentiment words also can represent the sentiment too.
- **Sarcasm**: a sentiment word with positive or negative review can switch sentiment if there is irony in the sentence for e.g. "Surely, I'm *happy* with my browser to collide right in the mid of my work".
- **Language**: Different languages have different words as a single word can change sentiment and meaning depending on the language. For example, the word "*sick*", have different meanings based on context, tone and language.

1.4 Measuring Accuracy of Sentiments

The accuracy of Sentiment Analysis can be calculated in many ways, but the most common approach is to score accuracy in comparison to a human. So any natural language processing engine that scores around 80% is consider great job with accuracy.

There are few major challenges for an engine analyzing text for sentiment. One of the biggest issues is that it has trouble understanding satire. Even humans have difficulty with someone who is being ironic. It is one of the most common mistakes a text analytics engine makes when trying to analyze text for sentiment.

Even humans have trouble, as they can analyze with 80% accuracy. Other problems are when words have multiple definitions. There are a few engines that use deep learning to help them understand context.[11]

1.5 Methods of Sentiment Analysis

Sentiment analysis can be categorized into four main groups: keyword spotting, lexical affinity, statistical methods, and concept-level techniques.

- Keyword spotting- in this classifying text based on the category presence of unambiguous words such as joyful, sad, scared, and tired of something.
- Lexical affinity- helps in detecting observable affecting words, but it also assigns subjective words a likely “affinity” for particular emotions or sentiments.
- Statistical methods leverage – it is based on elements of machine learning that includes latent semantic analysis and "bag of words".
- Concept-level leverage-it is based on essentials that consider knowledge representation forms such as ontologies and semantic networks and therefore are used for detecting semantics which provides meaning.[22]

1.6 Basic Components of an opinion

- It includes:
    - Opinion holder: Reviewer or an organization which holds a definite opinion on a particular or specific object or word.
    - Object: it is considered a word, phrase or sentence on which an specified opinion is expressed
    - Opinion: refers an attitude, review or appraisal on particular object from the side of an opinion holder [29].

1.7 Different level of Sentiment Analysis

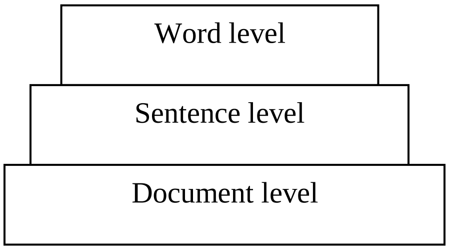


Figure 1: level of sentiment analysis

Different level for sentiment are as follows [28]:

1. Word level:

   It include following steps:

   a. Identifying and extracting various object features that have been given by opinion holder for e.g., reviewer.

   b. It determines whether the opinions on the particular feature of the object are positive, negative or neutral.

   c. Grouping the same features.

      i. Producing a summary of opinion of feature based on multiple reviews.

2. Sentence level:

   It includes following steps:

   a. Identify subjective or opinionated sentences

      i. Different Classes may be objective and subjective.

   b. Sentiment classification on each sentence.

      i. Different classes may be: positive, negative and neutral.

      ii. Assuming a sentence may contain only one opinion which may not be true in every case.

      iii. We can also consider clauses or phrases.

3. Document or review level: It includes:

   a. Sentiment classification of reviews on particular object.

   b. Different classes are positive, negative, and neutral

c. Assuming that each document or review mainly focuses on a particular object and contains opinion from a particular holder only.

1.8 Text Analysis process

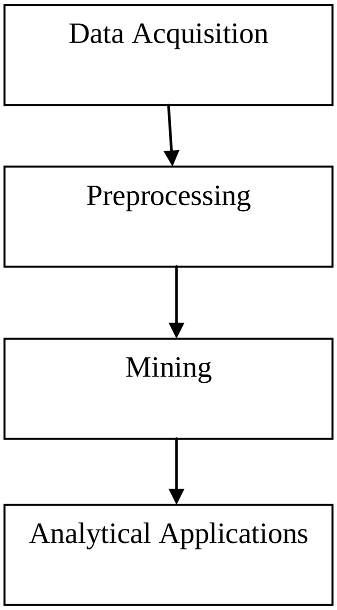


Figure 2 Process of analyzing Text

This process involves the following steps [29]:

1. **Data Acquisition**

In this data acquisition, data are gathered from different relevant sources such as web crawling, twitter tweets, online review, newsfeeds, document scanning etc.

I. **Web Crawling**: It is a procedure for automatically extracting web pages. It is consider an important component of search engine [18]. By using web page URL address, web crawler locates the all the pages. Includes following steps:

a. It starts from the page first
b. Read the whole content of the page

c. Parsing done on pages

d. Repeating the first step and circulate the procedure until all pages are not captured.

e. Extract the content of crawling pages with the help of Html parser.

II. **Document Scanning**: It is defined as a process to capture, store, and retrieve documents in spite of their original format, with the help of micrographics and electronic imaging i.e. scanning, OCR, ICR, etc. Document Scanning or Imaging is the process in which manuscript are copied and saved as digital imagery. These images are saved as PDF or TIFF files format.

2. **Preprocessing**: It is used to remove noisy, inconsistent and incomplete data. For doing the classification, Text preprocessing and feature extraction is a preliminary phase. Preprocessing involves 3 steps:

(i) Tokenization or segmentation: It is the process of splitting a string of written language into its words. Text data consists of block of characters referred to as tokens. So the documents are being separated as tokens and have been used for further processing.

(ii) Removal of stop words: Stop words are the words which are needed to be filtered i.e. may be before or after natural language processing. Stop words are words which contain little informational. Various tools specifically avoid to remove these stop words in order to support phrase search. Several collections of words can be chosen as stop words for any purpose. Some search engines, removes most of the common words which include lexical words such as "want" from a text in order to improve performance. Search engine or natural language processing may contain a variety of stop words. It includes English stop words such as “and”, “the”, “a”, “it”, “you”, “may”, “that”, “I”, “an”, “of” etc. which are considered as ‘functional words’ as they don’t have meaning. Researchers have shown that by removing stop words from the file, you can get the benefit of reduced index size without much affecting the accuracy of a user's. But care should be taken however to take into consideration the user's needs. Mostly, all search engines helps in eliminating the stop words from their indexes. With the help of eliminating stop words from the index, the index size can be reduced to about 33% for a word level index. While assessing the content of natural language processing, meaning of word can be conveyed more clearly by removing the functional word. If stop word removal is applied, stop list in one text file cannot be loaded [23].

(iii) Stemming: Stemming is the term which used to describe the process to reduce derived words to their origin word stem. Since 1960s, algorithms for stemming have been studied in the field of computer science [24]. Different Stemming methods are commonly referred as stemming algorithms or stemmers. For English, the stemmer example are that, it should identify the string “cats”, ”catty” as based on the root word “cat”, and also “walks”, “walked”, “walking” as based on the root word "walk".

a. Porter stemmer: It is defined as most primitive and best knowing algorithms used for stemming. This can be done heuristically by identifying word suffixes i.e. endings and strip them out, with few regularization at the end. It might collapse various sentiments, with the help of mapping two different words with different meaning into the same stemmed base. For e.g., Table1[26] consider class of positive and negative referred in the Harvard General Inquirer which maps two words with different sentiment:

Table 1: Mapping word by porter stemmer

Positive	Negative	Stemmed by porter
Common	Commoner	Common
Desirable	Desire	Desir
Affection	Affectation	Affect
Capitalize	Capital	Capit
Closeness	Close	Close

b. The Lancaster stemmer: It is a further broadly used algorithm for stemming. For sentiment analysis, Lancaster is more difficult than the Porter stemmer as it mix up more words of different word sentiment. Table 2[26] shows consider class of positive and negative referred in the Harvard General Inquirer which maps two words with different sentiment:

Table 2: Mapping word by Lancaster stemmer

Positive	Negative	Stemmed by Lancaster
Arbitration	Arbitrary	Arbit
Arbitrate	Arbitrary	Arbit
Call	Callous	Cal
Capitalize	Capital	Capit
Commitment	Commit	Commit

c. WordNet stemmer: It has high-precision functionality, but problem is that it is of limited use for sentiment analysis. For effecting changes, it may require a pair of word and part-of-speech tag, where the part-of-speech is considered as adjective -a, noun -n, adverb -r , or verb -v. When such pairs are given, it may collapse the tenses, aspect, and number marking [26].

Table 3 Mapping words by WordNet stemmer:

Word	Stem by WordNet
exclaims, verb	Exclaim
exclaimed, verb	Exclaim
proved, verb	Prove
proven, adjective	Proven

3. **Data mining**:

Applying different mining techniques to derive usefulness about stored information. Different mining approaches are classification, clustering, statistical analysis, natural language processing etc. In text analytics, mainly classification technique is used. Classification is a supervised learning method that helps in assigning a class label to an unclassified tuple according to an already classified instance set. Data classifying and identifying is all about to tag the data so it can be create quickly and efficiently.

But various organizations can gain from re-transforming their information, which helps in order to cut storage and backup costs, with increasing the speed of data searches. Classification can help an organization to meet authorized and regulatory requirements to retrieve specific information within a specific time period, and this is most important factor behind implementing various data classification technology.

1.9 Naïve Bayes Classifier

The Naive Bayes classifier is a simple one for calculating the probabilities that is based on Bayes theorem that have strong and naïve independency. It is considered as most useful classifying the text with number of applications such as detecting email spam, to sort personal, categorizing the document, detecting the language and detecting the sentiment. In spite of having the naïve design and simple assumptions, this classifier will perform better in various problem that are complex in real world.

Even though there are some other techniques such as random forests, Decision tree, rule based mining, Support Vector Machines etc, naïve bayes classifier works efficiently since it is not as much expensive in both CPU and memory usage and also needed a small dataset for training purpose. Also, the doing training with the Naive Bayes is consider small as compared to other classifier.

Different types of Naive Bayes Variation

There are several number of variations for Naive Bayes i.e. (i) Multinomial, (ii) the Binarized Multinomial and (iii) the Bernoulli.

i. Multinomial - It is used to refer when the number i.e. count of words means a lot in the technique for doing classification. Example for this type is when need for performing Topic Classification.

ii. The Binarized - It is referred when the occurrences of the words don't play an important role in classification. For example, in Sentiment Analysis, where it is no need as how much times we use the word "good" not considering the fact that he does.

iii. The Bernoulli - It is referred to use when the problem of not having the particular word matters a lot. For example, consider the concept in Spam or Adult Content Detection that gives a best quality of results.

Naive Bayes classifier use to calculate each word probability:

a. Estimating the probability $P(c)$ for each class c i.e. positive, negative and neutral by dividing the total number of words in documents in c by the total number of words in the corpus.
b. Estimating the probability distribution $P\ (w \mid c)$ for all words w and classes c where w is the number of tokens. This can be calculated by dividing the number of tokens of the words w in documents in c by the total number of words in c.
c. For scoring a document *doc* for class *c*, calculate

$$Score(doc, c) \stackrel{\text{def}}{=} P(c) * \pi \prod_{i=1}^{n} P(w_i \,|c)$$

d. To predict the most likely class label, then you can just pick the class c with the highest scoring value. To calculate the probability distribution, use equation:

$$P(c|doc) \stackrel{\text{def}}{=} \frac{score(doc, c)}{\sum_{c' \in C} score(doc, c')}$$

For testing purpose cross validation method is used. Cross-validation, it is also referred to as rotation estimation, is a model validating technique to assess how the results of a statistical analysis will simplify to an independent data set [25]. It is largely used in where the main goal is prediction, and estimating how accurately a predictive model will perform in real world. The main goal of cross validation is to define a dataset to "test" the data model in the training period i.e. in the validation dataset, so that to reduce the problem of over fitting, giving an insight on how the model will simplify to an independent dataset.

Two types of cross validation methods are there:

(i) Exhaustive: These are the methods which do learning and testing on all possible ways so as to divide the original sample into two parts i.e. (i) training and (ii) a validation set. It includes two type:

(a) Leave-p-out cross validation

(b) Leave one out cross validation.

(ii) Non-exhaustive: These are the methods that don't learning and testing on all possible of splitting the original sample. It also include two types:

(a) K-fold cross validation

(b) 2-fold cross validation

4. **Analytical Application**: It provides valuable things from text mining so that it can provide information that helps in improving decision and processes. It includes following ways such as sentiment analysis, document imaging, fraud analysis etc.

CHAPTER 2

LITERATURE REVIEW

Ana Mihanovic, Hrvoje Gabelica, Zivko Krstic. "***Big Data and Sentiment Analysis using Knime: Online Reviews versus Social media***", (**2014)** This paper analyze sentiment analysis on various gadgets in two different forms i.e. online review and Tweets. For dictionary making, it uses Knime tool in both forms. Online reviews have been crawled using Apache Nutch crawler while tweets were collected using Java package. As tweets are shorter so, number of tweets collection will be more compare to online reviews. Both tweets and online review are stored in HBase table on Apache Hadoop server. Data sets for online review are classified based on key, PID, Review Date, Review Text, Keyword, Language while Tweets are having attribute such as Key, UserScreenName, Creation Date, Text, Keyword, Language. This data is loaded into Knime. Dictionary build for online review can be easily categorized based on usage, price, quality, experience of user, Look and services provided by gadgets. Correct grading of phrases is to be done using grade scope. For tweets, Scoring is done based on polarity i.e. positive, negative or neutral. For tweets it will be difficult for scoring based on phrases as it is impossible to categorize them. They need more preprocessing and it uses frequency driven. So this paper concludes that Sentiment analyses on online review are less complicated and provides more detailed result as compared to tweets. Build a dictionary for tweets are complicated as it includes internet slang, sarcasm. So, social media are hard to analyze as they have their unique structure and grammar [1].

Mrs. R.Nithya, Dr. D.Maheshwari. "Sentiment Analysis on Unstructured Review", (2014) This paper represents Sentiment analysis that mainly on subjective and polarity detection. A proposed work include: (i) Feature Extract- Commonly, Sentiment analysis uses machine learning algorithm and a method to extract features from texts and then train the classifier. (ii) Preprocessing- stemming refers reducing words to their roots. Porter's stemming algorithm used for removing stop words. Mostly, adjective words have sentiment.

(iii) Product aspects- Textstat is a freely available that can be used for extracting pattern. (iv) Find polarity of opinionated sentence- here SentiStrength lexicon-based classifier used to detect sentiment strength. Here, 575 reviews have been taken from shopping sites. Tanagra1.4 tool used for data mining. Naïve bayes classification done through this tool based on each individual features such as display, accessories, battery life, weight and cost. Results shows that 'battery life' have most positive value so it improves branding and 'cost' have very low positive value that indicate seller to concentrate more on reputation and product quality [14].

Haruna Isah, Paul Trundle, Daniel Neagu. "*Social media Analysis for product Safety using Text mining and Sentiment Analysis*", (2014) This paper represents a framework to gather and analyze the view of users of drug and product by using text mining, sentiment analysis and machine learning. Proposed framework for processing view of customer on popular brand of drugs and cosmetic product include: (a) Text collection and cleaning, in this an API call for authenticating and extracting is invoked on Facebook Graph and Twitter APIs. Twitter API consists of REST and streaming APIs used to search and fetch tweets. Facebook Graph API used to fetch pages, update status and commenting on user experience. (b) Preprocessing phase- adapted bag of words representation technique. It includes the steps: remove delimiters, convert all words to lower case, remove numbers and stop words. (c) Sentiment analysis phase-It includes two approaches: (i) Lexicon based Sentiment-for improving classification it merges two lexicons application or domain specific lexicon and a generic English based lexicon. (ii) Machine Learning Based – a training dataset is needed and naïve bayes classifier is used due to its efficiency. (d) Evaluation phase-Contingency tables and truth tables used to represent the output of classifier. Two case studies were included: Sentiment analysis on facebook comments and Text mining and sentiment analysis of Twitter data. Future work includes clustering the tweets, temporal analysis, commenting, spamming and user sentiment by location [5].

Bogdon Batrinca, Philip C. Treleaven. "*Social Media Analytics: a survey of techniques, tools and Platform*", (2014) states an overview of software tool for social media, blogs, chats, newsfeeds etc. and how to use them for scraping, cleansing and analyzing. For scraping the social media it suggests the challenges such as Data cleansing, Data protection, Data analysis and Visualization and analytics Dashboard. This paper presents a survey on methodology of social media, data, providers and analytics techniques such as stream processing, sentimental analysis. An overview of different tools needed for social analysis purpose is also presented. There has been easy availability of APIs provided by Twitter, Facebook and News services which led to explosion of data services for the purpose of scraping and sentiment analysis [3].

Mohsen Farhadloo, Erik Rolland. "*Multi-class Sentiment analysis with clustering and score representation*", (2013) This paper represents improved method for aspect level sentiment analysis. It also proposes to use bag of noun instead of bag of words as to improve the clustering result for the aspect identification. By illustrating an example it is proven that bag of noun is able to find similar sentences using clustering algorithm as compared to bag of words. After identifying the aspects, the sentiment of each sentence need to be identified which contain one of those aspects. Here, SVM (Support Vector Machine) classifier has been used. Two type of representation i.e.(a) Bag of words (b) Score . In this paper, two experiments were done i.e. firstly, comparing bag of noun and bag of words clustering and secondly, classifying each sentence to positive, neutral or negative sentiment. Here, TripAdvisor.com used to create corpus. For 3-class classification, one-against-all scheme have been used. It use 5-fold cross validation. Results prove that clustering with bag of noun yields better and meaningful aspect than bag of words approach. By using 3-class sentiment analysis, we can improve the performance by 20% in terms of average f1-score [12].

V.K. Singh, R.Piryani, A. Uddin, P.Waila. "*Sentiment Analysis of Movie Reviews*", (2013) This paper represents an experiment on new kind of feature based for aspect-level sentiment analysis of movie review. Two algorithms were formulated using SentiWordNet library. They are: (a) Document-level Sentiment Classification. (b) Aspect- level Sentiment Analysis. In Document level, entire document have been classified into positive and negative class. First of all review should be applied to a POS tagger before it can be applied to SentiWordNet. For selected POS tag sentiment score of each extracted term is obtained from SentiWordNet library. Two linguistic feature selection schemes are used they are: (a) Extract only adjectives and adverbs, preceding the selected adjectives. (b) Extracting both adjective and verbs along with any adverbs. Results shows that 30% weight for verb scoring will produces best accuracy level. In aspect level, detailed analyses of review are considered. Steps involved are: (i) identify which aspect is needed to be analyzed. (ii) Locate the opinion context for that aspect in the review. (iii) Determine the sentiment polarity of views about an aspect. Dataset have been collected from the popular movie review website http://www.imdb.com. Out of 1000 movie review, SentiWordNet Scheme i.e. SWN AAC(adverb + adjective) have 82% of positive, SWN AAAVC(adverb + adjective and adverb + verb) combination have 82.9% as positive and Alchemy API have 73.4% as positive. Similarly for negative review 18% (SWN AAC), 17.1% (SWN AAAVC) and 26.6% (Alchemy API) [18].

Basant Agarwal, Narmita Mittal, Erik Cambria. "*Enhancing Sentiment Classification Performance using Bi-tagged Phrases*", (2013) This paper represents bi-tagged phrases has been used as features extraction in combination with unigram features for sentiment. Main objective is of designing a machine learning model which can classify a given movie review as positive and negative correctly. Here two types of features are extracted i.e. (i) unigram and (ii) bi-tagged phrase. A Bi-tagged phrase has been extracted using part of speech tag. Here, Dataset has been collected from Cornell Movie Review sites i.e. contain 1000 positive reviews and 1000 negative reviews for movie. Here SVM (support vector machine) and NB (naïve bayes) has been used for classifying the dataset into positive and negative sentiment polarity. Weka tool is used to implement these classifiers. Evaluate these classifier using 10 fold cross validation. Results shows that unigram feature performing individually can give better result compare to bi-gram and bi-tagged feature for both SVM and NB. Further, Bi-tagged phrases consider as features individually not performing well for sentiment classification. But if, bi-tagged phrases combined with the unigram feature can improve the performance of sentiment classification. The main drawback here will be that it is highly computationally expensive [2].

Mathew Smith, Christian Szongott, Benjamin Henne, Gabriele von Voigt. "***Big data privacy issues in public social media***", **(2013)** stated how the capabilities of mobile devices are affecting user's privacy. It also presents threat analysis which is classified into two categories i.e. home grown problem in which user upload without sufficient protection which affect user's own privacy. Second, someone is uploading the damage content of other people. It also include privacy analysis of different sites such as flicker, Face book, Picasa web and Google+, Locr and Instagram and PicPlz. It also presents an analysis of privacy related metadata, particularly location data contain in social media. As it concludes that 10% of all the photos taken by camera devices harm other people's privacy without knowing them. It also represents handling of the location based big data. It includes a concept of watchdog client a server side watchdog service. In it, concept to stay in control from social media uploaded by others has three types of services. Through the regular user account, Operated by the social networks and last one operated by third party i.e. stand alone service [11].

Javier Conejero, Peter Burnap, Omer Rana ,Jeffrey Morgan. "***Scaling Archived Social media Data Analysis using a hadoop cloud***", **(2013)** paper presents a COSMOS platform for sentiment and tension analysis on twitter dataset. Tool used for sentiment analysis is SeniStrength. To run application based on cloud environment, it uses virtualized Hadoop Clusters in Open Nebula. This system configuration used for performance aspects which shows how virtual server needs to be distributed as to reduce variability in the analysis performance. It also presents the architecture for data processing of COSMOS using Open Nebula and Hadoop. Processing performance comparison is done over Cardiff Cloud Tweet and UCLM Cloud Tweet which shows Cardiff Cloud have better performance due to its compute node has been more powerful than UCLM Cloud compute node. This paper involves future work to evaluate on bigger cloud environment and increase number of virtual cluster and Twitter message and improve performance with multiple concurrent users using the same cloud service. As using COSMOS we can add more nodes and workers to the problem and bring processing time down further [6].

Ya-Ting Chang, Shih-Wei Sun . "***A Real-time Interactive Visualization System for Knowledge Transfer from Social Media in a Big Data***"**(2013)** stated a proposed real-time interactive visualization system. This paper is contributed towards three objectives a) analyze and visualize system from a social media on a real time basis. b) Kinect camera and a mobile device is used to interact with the system. c) Knowledge has been transferred from social media in big data providing Geo-location of social media, suggesting the path or route and then generating images. A proposed real time system consists of three parts i.e.

i) Analysis and visualization –here data are collected from social media and sent for analysis purpose. For analyzing partnership of a social media it has used Node XL.

ii) A kinect camera is used to track a user. Movements of user are tracked and are displayed in Virtual Reality environment. For the identity purpose, once user login in proposed system their identity has been recorded by the user: fusing sensors on a mobile device and hand joint.

iii) Shows the relationship between users and related multimedia content at that time. This paper provides a comparison on social network relationship, location based service and data visualizing [20].

Nargiza Bekmamedova, Graeme Shanks. "*Social Media Analytics and Business Value: A Theoretical Framework and Case Study*", (2013) stated that Social Media Analytics (SMA) uses various methods to analyze different patterns in social media data that helps in making decisions. It proposes the SMA framework for understanding and explaining how this SMA brings value to an organization. Here case study has been considered of Bankco, a global financial institution with over 8 million customers in 10 countries. Analysis of SMA framework include: Awareness motivation, SMA resources: IT assets and SMA Capabilities and another one for dynamic capabilities and third one is Awareness benefits. It includes lessons which are:

i) SMA assets and technical capabilities may be outsourced
ii) Early adopters of SMA can gain competitive advantage
iii) Dynamic capabilities are important in using SMA
iv) Embedding SMA within organizational processes and routines is important.

Limitations include only one case study. There is a need to consider more case studies in order to refine and then develop the framework. Further step is to develop SMA framework using some expert interviews and more case studies of industries [15].

Simona Vinerean, Iuliana Cetina. "*The Effects of Social Media Marketing on Online Consumer Behavior*", (2013) stated to answer the question of how people interact on online and how they are engaged in online activities. Study include online activities of 236 Social media users, by identifying different types of users, a segmentation of these users and a linear model is designed to examine how different predictors are related to social networking sites that consider a positive impact on the respondents perception of online advertisements. This study can help to discover how to engage different types of audience in order to maximize the effect of the online marketing strategy. Limitation of study include with online questionnaires, which include unsystematic sampling procedures and low response rate. Future research can be measured based on demographic variables such as sex, age and social class [16].

V. S. Jagtap, Karishma Pawar. "*Analysis of different approaches to Sentence-Level Sentiment Classification*", (2013) This paper represents to analyze a solution for the sentiment based on fine-grained level, mainly on sentence level in which polarity is considered based on three categories i.e. positive, negative and neutral. Sentiment analysis is based on three levels i.e. document, sentence and feature level. In document level, it mainly determines the overall sentiment of the document based on the class attribute. In Sentence level, each sentence is considered as a separate unit and assuming that the sentence contains only one review. Mainly two task are there in sentence (a) subjective classification (b) sentiment classification. In feature level, it is used to produce feature-based review. This paper also presents a sentiment model which involves the following steps: (a) Review (b) Data preparation (c) review analysis (d) sentiment classification (e) results. It also describes application such as in (i) review related website. (ii) In sub-component technology (iii) In business and government Intelligence. Different approaches are mentioned i.e. machine learning, maximum entropy etc. This paper also presents two semantic approach i.e. (a) Corpus based (b) Dictionary based [19].

Jalaj S. Modha, Gayatri S. Pandi, Sandip J. Modha. "*Automatic Sentiment Analysis for Unstructured Data*", (2013) This paper presents a new approach for classifying and handling subjective as well as objective sentence of sentiment analysis. In proposed approach, it includes four steps: (a) first of all classify the sentence into two categories i.e. opinionated and non-opinionated, without regarding whether it is subjective or objective. (b) As having Opinionated sentence classify them as subjective or objective. (c) Classify subjective sentence into positive, negative and neutral. (d) classify objective sentences into positive, negative and neutral, providing semantic orientation for complex one. In this paper, Support Vector Machine (SVM), Naïve bayes, Bag of Words (BOW), POS (Part-of Speech), SentiwordNet, N-gram , Text mining and grammar rules technology used for classification and sentiment analysis. Here, domain used is Indian Political news article and preparing dictionary for this domain [7].

Kristin Glass and Richard Colbaugh. "*Estimating the sentiment of social media content for security informatics applications*", (2012) stated that inferring the sentiment of Social media data such as post is consider an important aspects for both Security analysts and technically. Sentiment analysis of social media data for the purpose of security is considered to be the modest level in terms of particular area of interest and the need for adapting new domain as a result of this can lead to the methods which perform poorly. This paper represents two computational methods for the sentiment expression in social media for the above challenges. It formulates text classification method and represents the model of data as a bipartite graph of document and words and only limited information is available regarding to any document or words. The first algorithm used here is a semi supervised sentiment classifier that combines knowledge of sentiment label of few document and word with unlabeled data i.e. online. The second algorithm represents a set of labeled document in the domain and used these data to calculate sentiment in the target domain [9].

Deptii D.Chaudhri, R.A. Deshmukh. "*Feature –based Approach for Review mining Using Appraisal Words*", (2012) This paper represents a framework which use appraisal words lexicon and feature extraction of the product for review categorization. The proposed framework will extract the review from amazon.com sites. It includes mainly four steps: (a) Web Crawling (b) Generating Structured review data (c) Sentiment analysis (d) Review categorization. Feature extraction can be done through sentence splitting, part of speech tagging. Here, features are being expressed using noun and noun phrases and then their term frequencies were calculated. Here, features include 'touch screen', 'voice quality' etc. Aim is to find the sentiment about the product being expressed base on the product feature. Here, Apriori algorithm has been used. Association mining is used on the set of noun and noun phrase. For evaluating purpose, the proposed system is being evaluated from various perspectives such as effectiveness of the feature extraction and accuracy. For experiment 100 reviews for 5 products are considered. For evaluating the performance, precision and recall of these products are considered [4].

SitaramAsur, Bernardo A.Huberma, "*Predicting the Future with Social Media*", (2012) stated that how the content of social media can be used to forecast the outcome. Chatter has been used from twitter to forecast box-office revenue for movies. Dataset are obtained by crawling hourly feeding data from twitter by using Twitter Search Api. Extracting 2.89 million tweets referring to 24 different movies released over a period of three months. This paper presents a linear regression model to predict box-office revenue of movie in advance of their release. It also shows the result in terms of accuracy and also there is a strong correlation between attention given to an upcoming movie and its ranking in future. The work here represents how social media expresses a collective wisdom can yield an extremely powerful and accurate indicator of future outcomes [17].

Jiao Wu, Weihua Gao, Bin Zhang, Yi Hu, Jinsong Liu. "*Online Web Sentiment Analysis on Campus Network*" ,(2011) This paper presents that multiple participants uses Web page to discuss one topic in parallel. Mostly, content include opinion, attitude, feeling rather than just facts. So, the proposed system classifies the reviews and calculates their sentiment rate by analysis. The design adopted by the system consists of Front-monitor Node which are distributive structure, Spider Nodes and Back Analysis Nodes which are cluster nodes and dual controller nodes are the hot-spare mode. Here, server cluster node runs on virtual platform. There are 4 Intel Xeon 2.27GHz CPU, 4G *32 DDr Dram, 8 Gigabit Network Interface Card and 10 Tb Optical Storage array. Here, an Online Web Sentiment analysis has been presented which is based on cluster service so that it can achieve high performance and high availability. Results show that same subjective word can go through different ISR (Integrated Sentiment Rate) by Basic Sentiment rate, Internet Topic Rate, Manual Influence rate and Campus User rate [10].

Ms. K. Mouthami, Ms. K.Nirmala Devi, Dr. V.Murali Bhaskaran. "*Sentiment Analysis and Classification based on Textual Reviews*", (2010) This paper represents a new algorithm i.e. Sentiment Fuzzy Classification with part of speech tag used to improve accuracy of the classification of movie review dataset. It uses fuzzy set because it represents the interior fuzziness in sentiment analysis. Steps include Text Preprocessing, Transformation, Feature Selection, Classification and Evaluation. Text preprocessing divided into two part: (i) Tokenization (ii) Removal of Stop words. Feature Selection involves two-step process: (a) Identifying the parts of the document which forms positive and negative sentiments. (b) Join these parts in such a way that they fall into one of these two polar categories. This paper refers to document level sentiment classification. Dataset are collected by Cornell movie review corpora. Sentiment Fuzzy Classification algorithm is proposed to improve classification accuracy for the movie review dataset [14].

Lingyan Ji, Hanxiao Shi, Mengli, Mengxia Cai, Peiqi Feng. "*Opinion Mining of Product reviews based on semantic role labeling*", (2010) This paper represents a sentiment analysis and retrieving system has been proposed which helps to mine useful knowledge from product reviews. Most complicated problem today, of textual analysis is of various domestic languages other than English. The proposed system consists of three parts: (a) Web Crawling-automatically extracting web pages. To extract, HTML parser using semi-automatic method is used. (b) Opinion processing- it includes establishing of polarity dictionary and analysis of emotion or feature tendency. (c) Interface Display- it includes: (i) Visualizing comparison between characteristics of product. Here, camera has been taken and for analysis purposes consider 100 reviews. It shows that the consumer being satisfied with the camera's price but it may have a neutral view towards the camera's picture. (ii) Visualizing opinion Contrast- it shows that the consumer prefer the picture of the camera. In future scope, System will include three objectives i.e. (i) development of technology (ii) service application and (iii) forecasting the trend [10].

CHAPTER-3
PRESENT WORK

3.1 Problem Formulation

Different types of data are generated from different Social media groups that need to be organized and to monitor people’s attitude towards products, gadgets, movie review etc. This database is collected from different social media sites for example Twitter, Facebook, Online review, shopping sites etc. Text analytics and Sentiment analysis can help to develop valuable business insights from text based contents that may be in the form of word documents, tweets, comments and news that related to Social media. The foremost reason of Sentiment analysis is so complex is that words often take different meanings and are associated with different emotions depending on the domain in which they are being used. Dataset is analyzed by using the weka tool. The hidden relationship has to be extracted from this type of database using different mining approaches in Weka tool. Dictionary building for detailed sentiment analysis implies making an initial list of adjectives and nouns which are normally used when describing a specific movie review. Phrases and terms are extracted from this relational dataset and their meaning has been added to dictionary for next generation analysis. In tweets, informal and shortcuts has been used for explaining terms or views and this is done with the help of sentiments analysis is not an easy process. To reduce this, data mining approaches has been used for extraction of features from these datasets.

3.2 Objective

- To implement data extraction using Weka tool.
- To classify data using naïve bayes multinomial to determine their positive, negative and neutral polarity distribution.
- To construct combined dictionary from online review and Twitter dataset keywords i.e. from movie reviews.
- To analyze parameter for performance analysis.
- To remove ambiguous data.

3.3 Methodology

In this proposed system, Weka an open source data mining tool has been used so as to perform sentiment classification on movie review dataset. Here, goal is to classify dataset into positive and negative and form the combined dictionary of Twitter dataset and online review dataset. Main steps are:

1) Generating Dataset

Two dataset were collected firstly, from Twitter tweets and secondly, from Online review Dataset. The online review dataset consists of around 800 user's review archived on the IMDB (Internet Movie Database) portal. And for, Twitter dataset around 1000 review were collected and each review were formatted according to .arff file where review text and class label are only two attributes. Class label represent the overall user opinion. Here, we set simple rules for scaling the user review. For dataset, a user rating greater than 6 is considered as positive, between 4 to 6 considered as neutral and less than 4 considered as negative.

2) Preprocessing

For doing the classification, Text preprocessing and feature extraction is a preliminary phase. Preprocessing involves 3 steps:

(i) Word parsing and tokenization: In this phase, each user review splits into words of any natural processing language. As movie review contains block of character which are referred to as token.

(ii) Removal of stop words: Stop words are the words that contain little information so needed to be removed. As by removing them, performance increases. Here, we made a list of around 320 words and created a text file for it. So, at the time of preprocessing we have concluded this stop word so all the words are removed from our dataset i.e. filtered.

(iii) Stemming: It is defined as a process to reduce the derived words to their original word stem. For example, “talked”, “talking”, “talks” as based on the root word “talk”. We have used Snowball stemmer to reduce the derived word to their origin.

3) Classification

Classification is a supervised learning method that helps in assigning a class label to an unclassified tuple according to an already classified instance set. Here, naïve bayes multinomial classifier has been used. Quality measure will be considered on the basis of percentage of correctly classified instances. For the validation phase, we use 10-fold cross validation method. Naïve bayes multinomial helps in generating dictionary and frequent set. It counts the occurrences of words in whole dataset and forms a dictionary of some most frequently occurring words.

Attributes are referred as text positions, values are referred as words.

$$\begin{aligned} c_{NB} &= \underset{c_j \in C}{\operatorname{argmax}}\ P(c_j)\prod_i P(x_i \mid c_j) \\ &= \underset{c_j \in C}{\operatorname{argmax}}\ P(c_j)P(x_1 = \text{"our"} \mid c_j)\cdots P(x_n = \text{"text"} \mid c_j) \end{aligned}$$

From training the corpus huge amount of data, extract Vocabulary. Calculate the probability of $P(c_j)$ and $P(x_k \mid c_j)$ terms. For each c_j in C do. $docs_j$ which is referred as the total number of documents for which the target class is c_j

$$P(c_j) \leftarrow \frac{|docs_j|}{|\text{total\# documents}|}$$

$Text_j$ it is referred as single document containing all $docs_j$ .For each word x_k in *Vocabulary*. $n_k \leftarrow$ number of occurrences of x_k in $Text_j$

$$P(x_k \mid c_j) \leftarrow \frac{n_k + \alpha}{n + \alpha \,|Vocabulary|}$$

- positions ← all word positions in current document which contain tokens found in *Vocabulary*

Return c_{NB}.

4) Parameter Evaluation

Now for evaluating the result, different parameter are to be calculated. True positive, True negative, False positive and False negative are used for comparing the class label that have been assigned to a document by the classifier with the classes the item actually belongs [19].

a) Accuracy: It is measured as the proportion of correctly classified instances to the total number of instances being evaluated. Classification performance being evaluated by using this parameter [19].

$$\frac{True\ positive\ \ + True\ negative}{True\ positive\ \ + True\ negative\ \ + False\ positive\ \ + False\ negative}$$

Where True positive – that are truly classified as positive.

False positive- not labeled by the classifier as positive but should be

True negative- that are truly classified as negative

False negative- not labeled by the classifier as negative but should be

b) Precision: It is widely used in evaluating the performance in different field such as text mining, information retrieval. Precision is also referred to measure the exactness. It is defined as ratio of the number of correctly labeled as positive to the total number that has been classified as positive [19].

$$precision = \frac{true\ positive}{true\ positive + false\ positive}$$

c) Recall: It is also used in evaluating the performance for text mining and information retrieval. It is also used to measure the completeness of the model. It is defined as the ratio of the number of correctly labeled as positive to the total number that are truly positive [19].

$$recall = \frac{true\ positive}{true\ positive\ \ + false\ negative}$$

d) F-measure: It is referred as the harmonic mean of precision and recall. It helps to give score needed to balance between precision and recall. It combines two of them into one for the convenience as it might optimize the system so that it can favor one of them [19].

$$f = \frac{2 \times precision \times recall}{precision + recall}$$

5) Combined dictionary

Combined word of twitter dataset and online review dataset forms a dictionary. As after classifying each word probability as positive, negative and neutral. Compare the probability for each word and categorize each word into three different dictionaries based on highest polarity of each word. Dataset is used for further evolution of words depending on their uses in daily life as adjectives or nouns in the social media data.

Proposed System Flow work

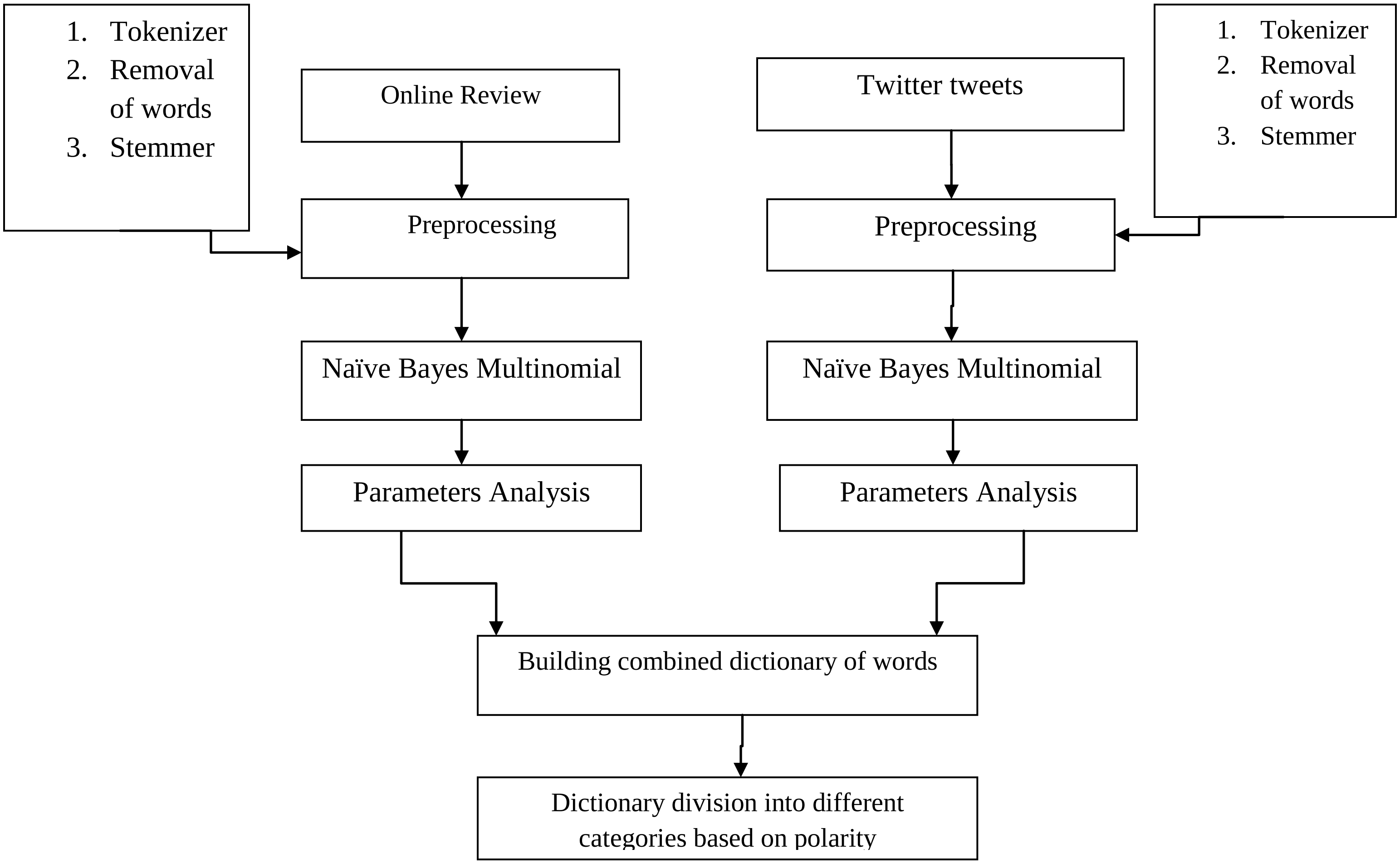


Figure 3: Flowchart

CHAPTER-4

RESULTS AND DISCUSSION

Graphical representation of GUI (Graphical User Interface) for this system is:

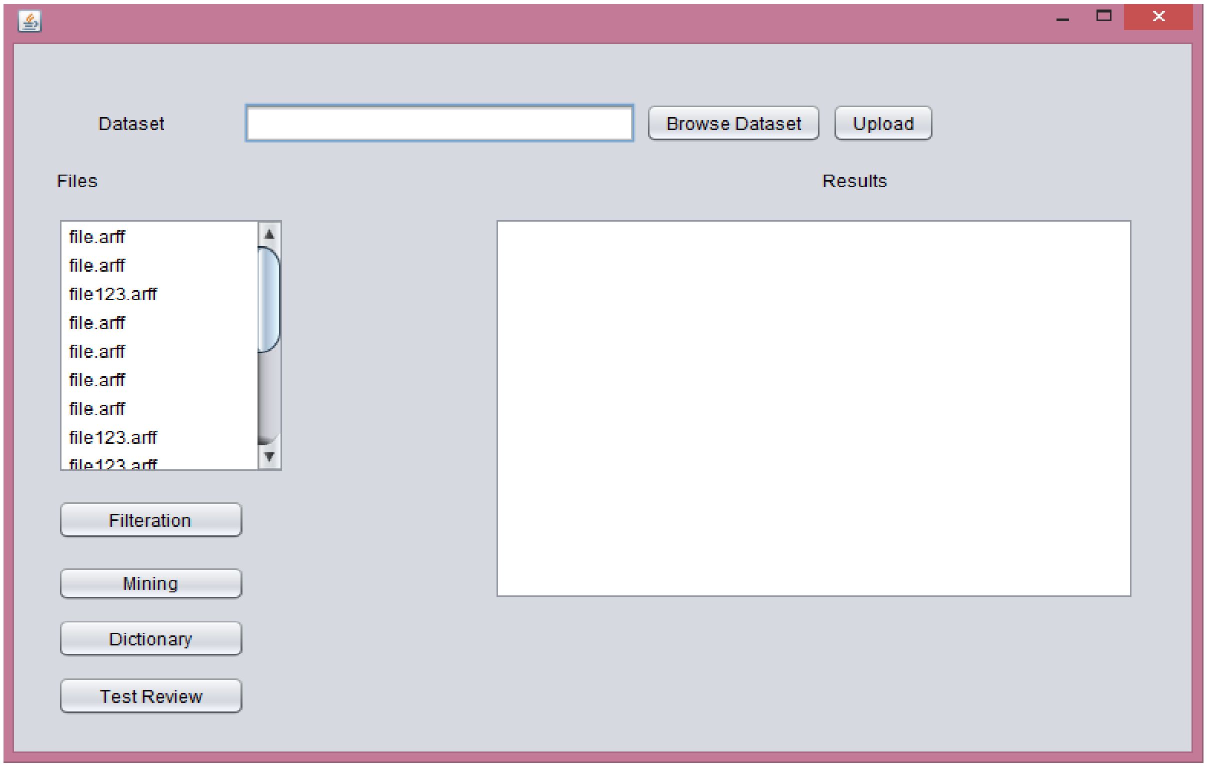
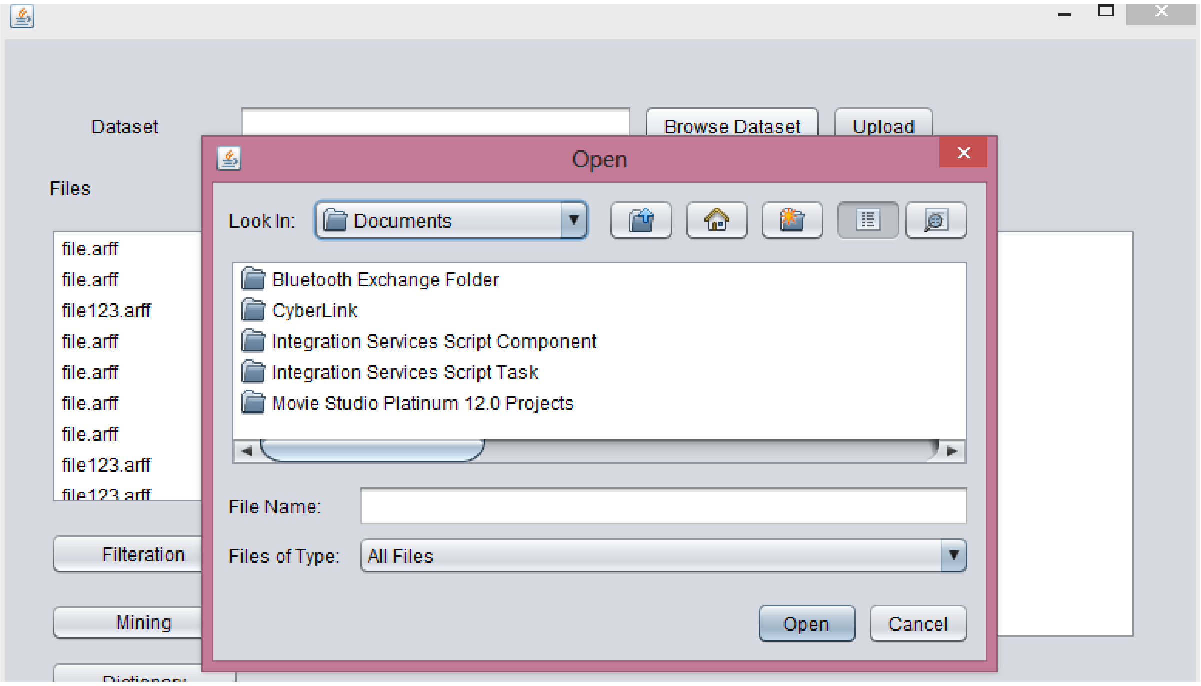


Figure 4: Snapshot of GUI

We have build this GUI using Net Beans IDE 7.1.2. For browsing the dataset in the GUI.



file.arff
file.arff
file123.arff
file.arff
file.arff
file.arff
file.arff
file123.arff
file123.arff

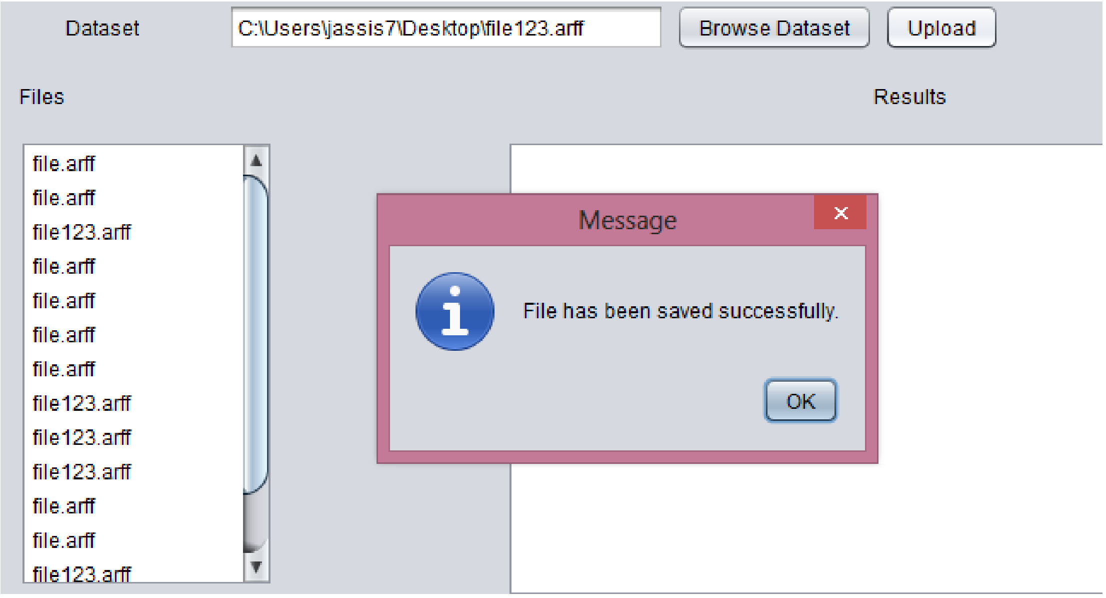
Look In: Documents

Bluetooth Exchange Folder
CyberLink
Integration Services Script Component
Integration Services Script Task
Movie Studio Platinum 12.0 Projects

Files of Type: All Files

Figure 5 Snapshot of Browsing the Dataset

We have used the review of online review and twitter tweets of different movies for analyzing. Table shows the number of collection of reviews for both online review and twitter tweets. As after browsing the dataset upload it.

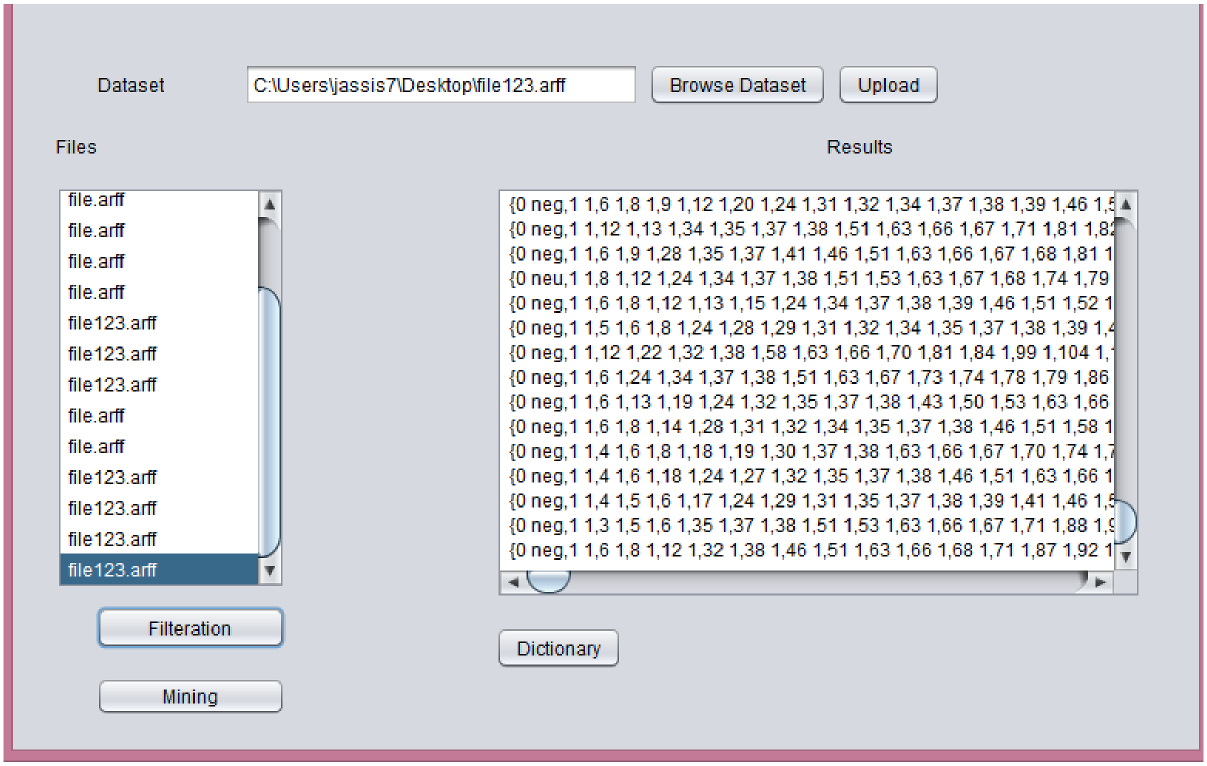


Dataset: C:\Users\jassis7\Desktop\file123.arff

file.arff
file.arff
file123.arff
file.arff
file.arff
file.arff
file.arff
file123.arff
file123.arff
file123.arff
file.arff
file.arff
file123.arff

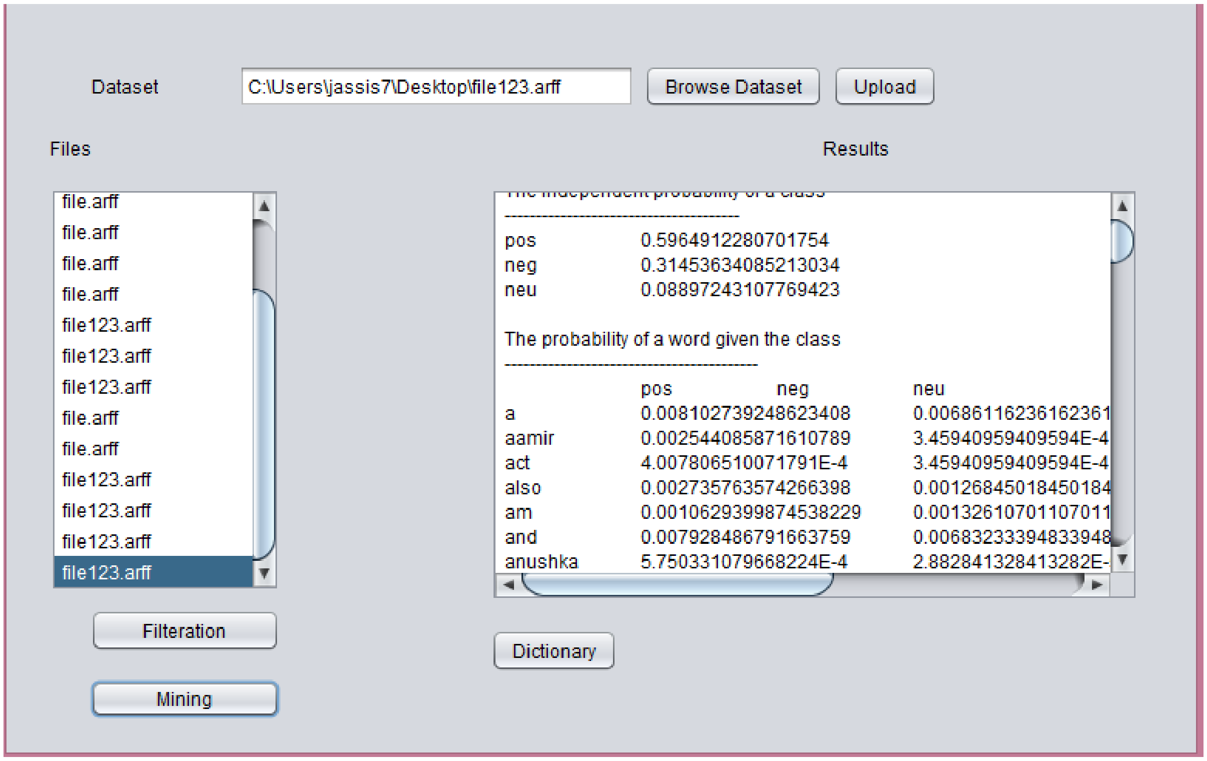
File has been saved successfully.

Figure 6 Snapshot of Uploading the Dataset



```
{0 neg,1 1,6 1,8 1,9 1,12 1,20 1,24 1,31 1,32 1,34 1,37 1,38 1,39 1,46 1,5
{0 neg,1 1,12 1,13 1,34 1,35 1,37 1,38 1,51 1,63 1,66 1,67 1,71 1,81 1,82
{0 neg,1 1,6 1,9 1,28 1,35 1,37 1,41 1,46 1,51 1,63 1,66 1,67 1,68 1,81 1
{0 neu,1 1,8 1,12 1,24 1,34 1,37 1,38 1,51 1,53 1,63 1,67 1,68 1,74 1,79
{0 neg,1 1,6 1,8 1,12 1,13 1,15 1,24 1,34 1,37 1,38 1,39 1,46 1,51 1,52 1
{0 neg,1 1,5 1,6 1,8 1,24 1,28 1,29 1,31 1,32 1,34 1,35 1,37 1,38 1,39 1,4
{0 neg,1 1,12 1,22 1,32 1,38 1,58 1,63 1,66 1,70 1,81 1,84 1,99 1,104 1,1
{0 neg,1 1,6 1,24 1,34 1,37 1,38 1,51 1,63 1,67 1,73 1,74 1,78 1,79 1,86
{0 neg,1 1,6 1,13 1,19 1,24 1,32 1,35 1,37 1,38 1,43 1,50 1,53 1,63 1,66
{0 neg,1 1,6 1,8 1,14 1,28 1,31 1,32 1,34 1,35 1,37 1,38 1,46 1,51 1,58 1
{0 neg,1 1,4 1,6 1,8 1,18 1,19 1,30 1,37 1,38 1,63 1,66 1,67 1,70 1,74 1,7
{0 neg,1 1,4 1,6 1,18 1,24 1,27 1,32 1,35 1,37 1,38 1,46 1,51 1,63 1,66 1
{0 neg,1 1,4 1,5 1,6 1,17 1,24 1,29 1,31 1,35 1,37 1,38 1,39 1,41 1,46 1,5
{0 neg,1 1,3 1,5 1,6 1,35 1,37 1,38 1,51 1,53 1,63 1,66 1,67 1,71 1,88 1,9
{0 neg,1 1,6 1,8 1,12 1,32 1,38 1,46 1,51 1,63 1,66 1,68 1,71 1,87 1,92 1
```

Figure 7: Snapshot of Result after filtering the dataset.



```
The independent probability of a class
------------------------------------
pos            0.5964912280701754
neg            0.3145363408521303 4
neu            0.08897243107769423

The probability of a word given the class
---------------------------------------
               pos                  neg                  neu
a              0.008102739248623408                      0.006861162361623 61
aamir          0.002544085871610789                      3.45940959409594E-4
act            4.00780651007179 1E-4                     3.45940959409594E-4
also           0.002735763574266398                      0.001268450184501 84
am             0.0010629399874538229                     0.001326107011070 11
and            0.007928486791663759                      0.006832333948339 48
anushka        5.750331079668224E-4                      2.882841328413282E-
```

Figure 8: Snapshot of showing the probability of each word.

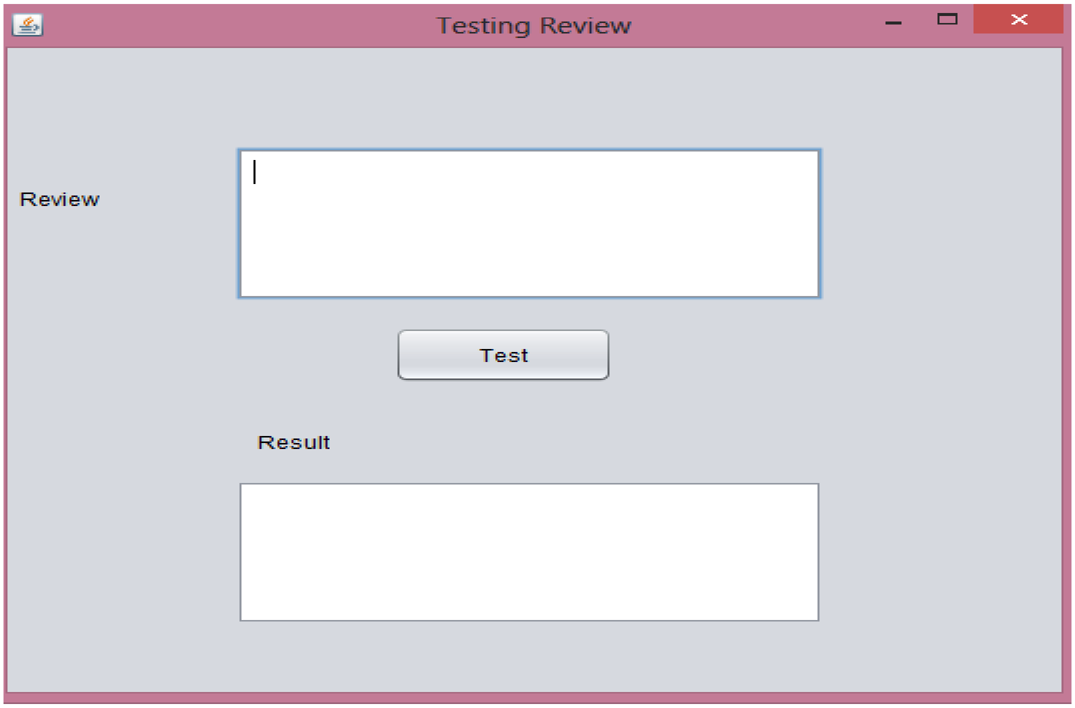
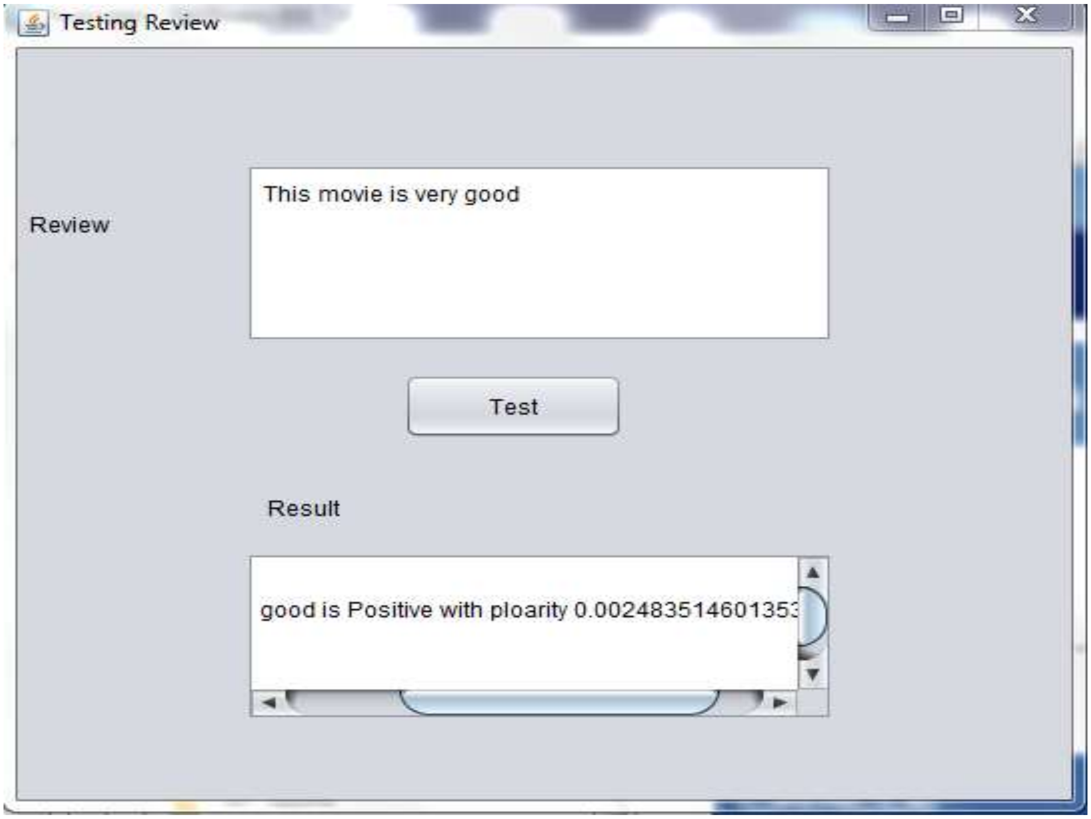


Figure 9 Testing to view the review



This movie is very good

```
good is Positive with ploarity 0.002483514601353
```

Figure 10 Result show the dictionary with polarity

Table 4 Collected Review of Online and Twitter Tweets

	Total number of review retrieved
Online Review	790
Twitter tweets	1149

After collecting dataset, we have preprocessed data. In preprocessing, three main steps are done as that we can the polarity of each word.

Three steps are Sentence splitting using string to word vector, Stop word removal and Stemming by Snowball stemmer.

Table 5 shows the results for distribution of sentiment score based on preprocessing result i.e. classified the dataset into three categories i.e. positive, negative and neutral.

Table 5 Distribution of Sentiment Score for Reviews

Score	**Reviews**	
	Online Review	Twitter tweets
Positive	475	499
Negative	238	494
Neutral	82	157

Figure 3 shows the positive, negative and neutral polarity for each word. Visualization of this can be done through weka tool.

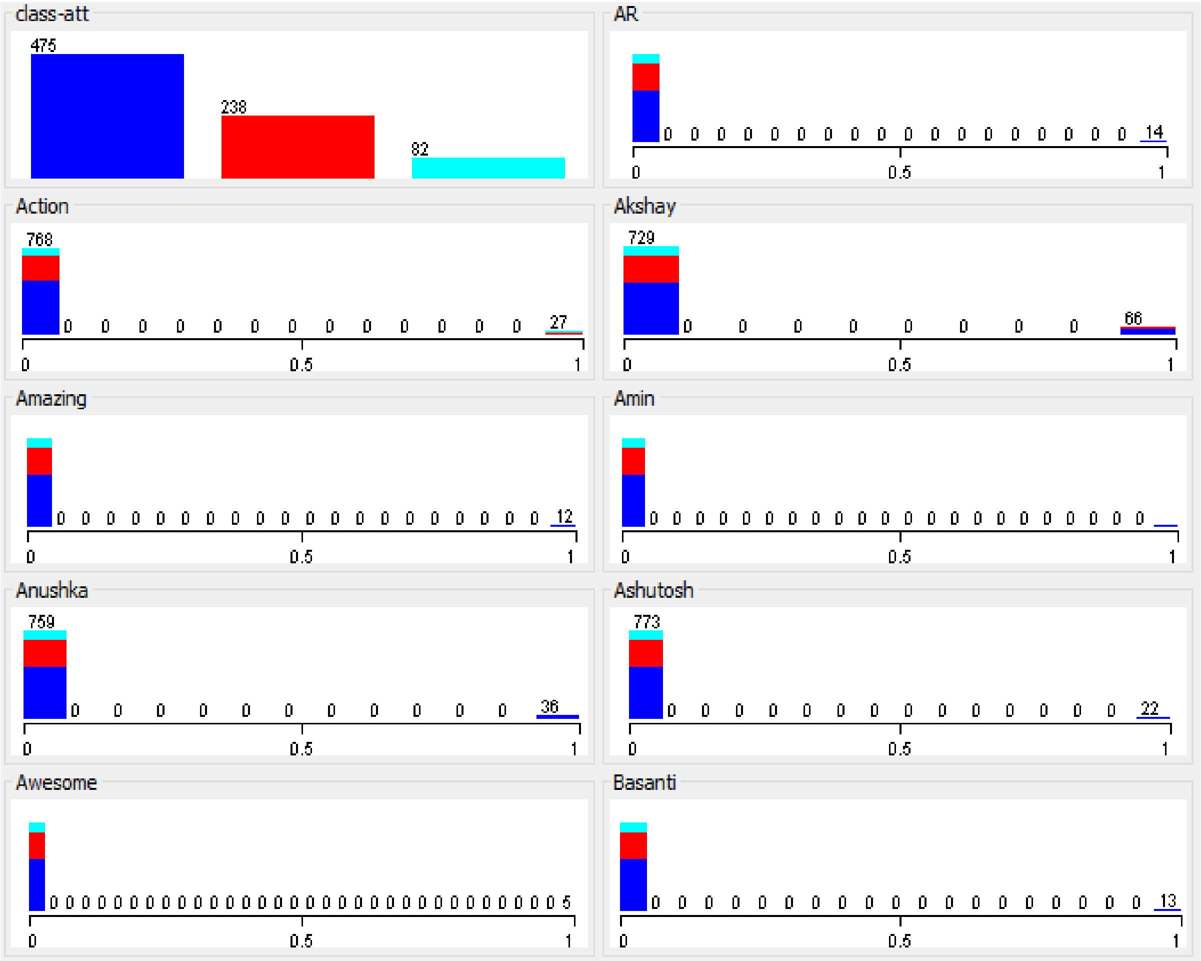


Figure 11 Visualization of Word Polarity

Now using Naïve Bayes classifier we have classified each word probability which is made through our Weka tool.

In online review dataset, 748 instances are correctly classified and for Twitter tweets, 974 instances are correctly classified.

Table 6 Performance of the naïve bayes classifier

	Online review	Twitter tweets
Correctly classified instances	748	974
Incorrectly classified instances	47	176

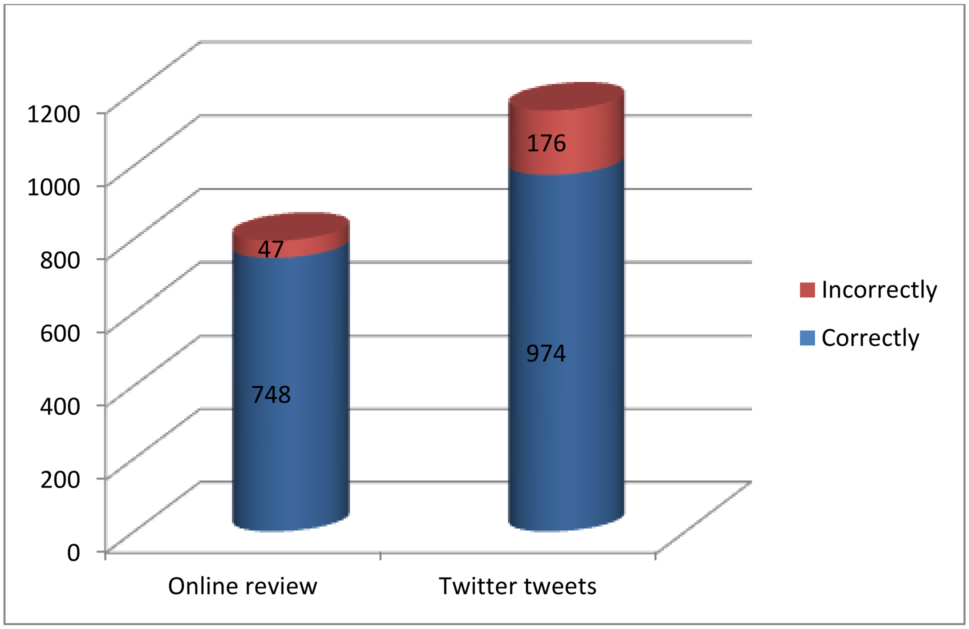


Figure 12: Performance of Online review and twitter tweets

Accuracy for online review of movie is around 94.08% and for twitter tweets is around 84.695%. This is due to as online review are written in detail and clearly viewed while tweets are short and have internet slang etc. so performance of online review is comparatively better than twitter tweets.

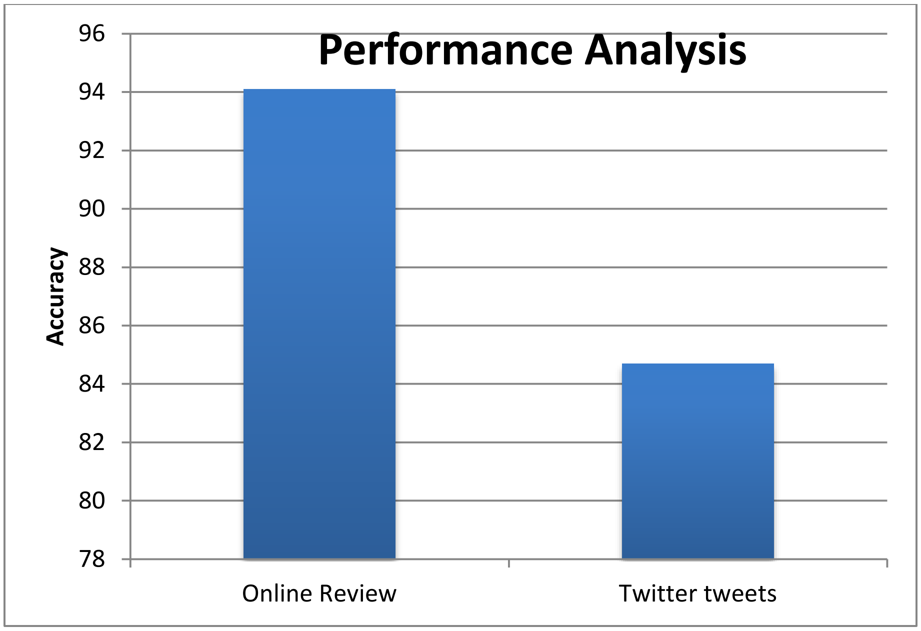


Figure 13 Performance analysis based on accuracy

Table 7 Detailed comparison of accuracy by the class attribute

	Online review	Twitter tweets
Kappa Statistic	0.8892	0.7479
Mean absolute error	0.04	0.125
Root mean squared error	0.1869	0.273
Relative absolute error	0.11	0.308
Root relative squared error	0.439416	0.606

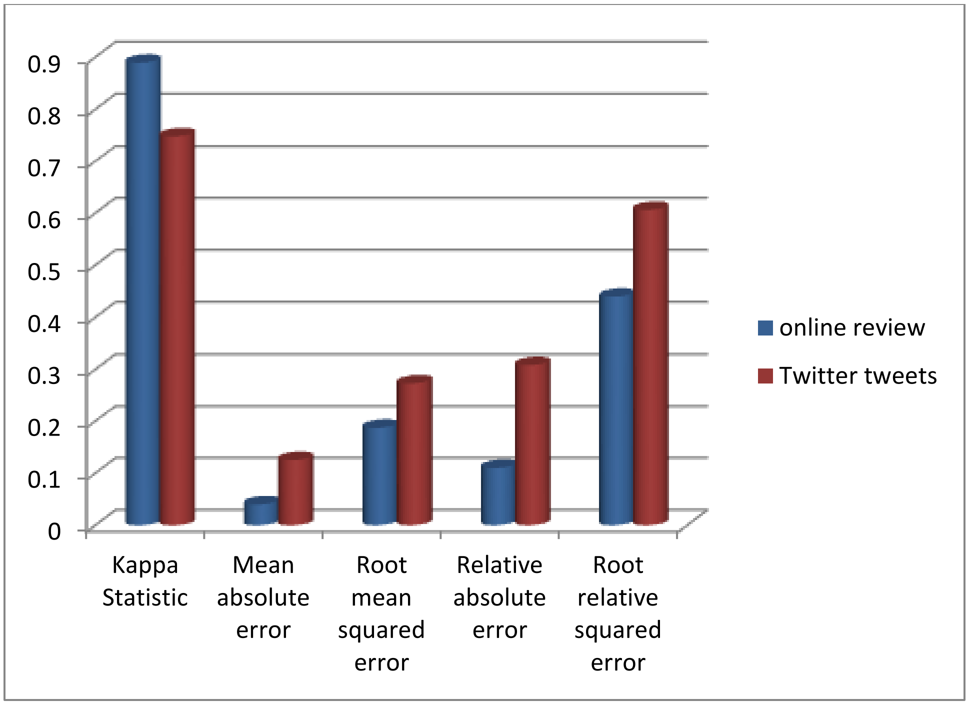


Figure 14 Evaluation of Parameter

For both online review and twitter tweets of movie detailed accuracy were calculated which includes parameter as TP rate, FP rate, Precision, Recall, F-measure and ROC area.

Table 8 Detailed Accuracy for Online Review

	TP rate	FP rate	Precision	Recall	F-measure	ROC Area
Positive	0.983	0.038	0.975	0.983	0.979	0.996
Negative	0.971	0.061	0.872	0.971	0.918	0.991
Neutral	0.61	0.001	0.98	0.61	0.752	0.968
Weighted Avg.	0.941	0.041	0.945	0.941	0.937	0.992

Table 9 Detailed Accuracy for Twitter tweets

	TP rate	FP rate	Precision	Recall	F-measure	ROC Area
Positive	0.866	0.111	0.857	0.866	0.861	0.959
Negative	0.868	0.101	0.867	0.868	0.868	0.958
Neutral	0.72	0.038	0.748	0.72	0.734	0.948
Weighted Avg.	0.847	0.096	0.846	0.847	0.847	0.957

Table 10 Online review confusion matrix

		Predicted Class		
Known Class		Positive	negative	Neutral
	Positive	467	8	0
	Negative	6	231	1
	Neutral	6	26	50

Table 11 Twitter dataset confusion matrix

		Predicted Class		
Known Class		Positive	negative	Neutral
	Positive	432	48	19
	Negative	46	429	19
	Neutral	46	18	113

CHAPTER-5

CONCLUSION AND FUTURE SCOPE

Social media Monitoring has been growing very rapidly so there is a need for various organizations to analyze customer behavior or attitude of particular product or any movie reviw. So, the concepts of sentiment analysis have been introduced. Text analytics and sentiment analysis can help organization to derive valuable business insights. Attitude can be calculated based on polarity check. Sentiment analysis on Online review are done by forming dictionary which shows that it is easier to build dictionary on phrases but complex in case of Twitter as tweets consist of short hands as online review were written in more clear way as compared to Tweets. So, form hidden relationship between different keywords and a dictionary of the words on the basis of different categories of comments & tweets.

Future work include to determine their features for the movie in detail i.e. make polarity check on different features such as actors, directors, scripts, music etc. and make the dictionary for them.

CHAPTER-6

REFERENCES

I. Research Papers

[1] Ana Mihanovic, Hrvoje Gabelica, ZivkoKrstic (2014) “Big Data and Sentiment Analysis using Knime: Online Reviews Vs. Social Media”, MIPRO Opatija, Croatia.

[2] Basant Agarwal, Narmita Mittal, Erik Cambria. (2013) “Enhancing Sentiment Classification Performance using Bi-tagged Phrases”, 13th International Conference on Data Mining Workshops, IEEE.

[3] Bogdon Batrinca, Philip C. Treleaven (2014) “Social media analytics: a survey of techniques, tools and platform” Department of Computer Science, Gower Street, London, UK published in Springer.

[4] Deptii D.Chaudhri, R.A. Deshmukh. (2012) “Feature –based Approach for Review mining Using Appraisal Words”, Department of Post Graduate Computer Engineering, Pune, India, IEEE.

[5] Haruna Isah, Paul Trundle, Daniel Neagu. (2014)”Social media Analysis for product Safety using Text mining and Sentiment Analysis”, Artificial Intelligence Research Group, University of Bradford, UK, IEEE.

[6] Jalaj S. Modha, Gayatri S. Pandi, Sandip J. Modha. (2013) “Automatic Sentiment Analysis for Unstructured Data”, published in IJARCSSE.

[7] Javier Conejero, Peter Burnap, Omer Rana, Jeffery Morgan (2013) “Scaling Archied Social Media Data Analysis Using a Hadoop Cloud” sixth international conference on cloud computing, IEEE.

[8] Jiao Wu, Bin Zhang, Weihua Gao, Yi Hu, Jinsong Liu (2011),"Online Web Sentiment Analysis on Campus Network", 4th International Symposium on Computational Intelligence and Design, IEEE.

[9] Kristin Glass and Richard Colbaugh (2012) "Estimating the sentiment of social media content for security informatics applications", Institute for Complex Additive System Analysis, Socorro, USA, Springer Journal.

[10] Lingyan Ji, Hanxiao Shi, Mengli, Mengxia Cai, Peiqi Feng. (2010) "Opinion Mining of Product reviews based on semantic role labeling", 5th International Conference on Computer Science and Education, IEEE.

[11] Matthew Smith, Christian Szongott, Benjamin Henne, Gabriele von Voigt (2013) "Big Data Privacy Issues in Public Social Media", Distributed Computing & Security Group, Leibniz Universitat Hannover, Thailand, Germany IEEE.

[12] Mohsen Farhadloo, Erik Rolland. (2013) "Multi-class Sentiment analysis with clustering and score representation", 13th International Conference on Data mining Worshops, IEEE.

[13] Mrs. R.Nithya, Dr. D.Maheshwari. (2014) "Sentiment Analysis on Unstructured Review",International Conference on Intelligent Computing Application, IEEE.

[14] Ms. K. Mouthami, Ms. K.Nirmala Devi, Dr. V.Murali Bhaskaran. (2010) "Sentiment Analysis and Classification based on Textual Reviews", Dept of CSE, Tamil Nadu, IEEE.

[15] Nargiza Bekmamedova, Graeme Shanks (2013) "Social Media Analytics and Business Value: A Theoretical Framework and Case Study", Department of Computing and Information Systems, University of Melbourne.

[16] Simona Vinerean, Iuliana Cetina (2013) "The Effects of Social Media Marketing on Online Consumer Behavior", International Journal of Business and Management; Vol. 8, No. 14.

[17] SitaramAsur, Bernardo A.Huberma (2012) “Predicting the Future with Social Media”, Social Computing Lab, HP Labs, Palo Alto, California.

[18] V.K. Singh, R.Piryani, A. Uddin, P.Waila. (2013) “Sentiment Analysis of Movie Reviews”, Department of Computer Science, New Delhi, India, Published in IEEE.

[19] V. S. Jagtap, Karishma Pawar. (2013) “Analysis of different approaches to Sentence-Level Sentiment Classification”, published in IJSET.

[20] Ya-Ting Chang, Shih-Wei Sun (2013) “A Real time Interactive Visualization System for Knowledge Transfer from Social Media in a Big Data”, Center for Art and Technology, Taipei National University of the Arts, Taipei, Taiwan, IEEE.

II.Websites

[21] https://semantria.com/sentiment-analysis

[22] http://en.wikipedia.org/wiki/Sentiment_analysis

[23] http://en.wikipedia.org/wiki/Stop_words

[24] http://en.wikipedia.org/wiki/Stemming

[25] http://en.wikipedia.org/wiki/Cross-validation_(statistics)

[26] http://sentiment.christopherpotts.net/

[27]. http://www.parliament.uk/briefing-papers/post-pn-460.pdf

[28] www.cse.unt.edu/~tarau/.../NLP/15SubjectivityandSentimentAnalysis.ppt

[29] www.decideo.fr/bruley/docs/2___sentiment_a_v0.ppt

CHAPTER-7

APPENDIX

Abbreviations

URL- Uniform Resource Locator

OCR – Optical Character Recognition

ICR- Intelligent Character Recognition

TIFF- Tagged Image File Format

KNIME- Konstanz Information Miner

WEKA - Waikato Environment for Knowledge Analysis

REST – Representational State transfer

BOW- Bag of Words

BAN –Bag of Noun

POS – Part of Speech

SVM – Support Vector Machine

SMA – Social Media Analytics