UNIVERSITY OF MEMPHIS

SENTIMENTAL ANALYSIS

Venkata Krishna Chikkala

Swaroop Sudheer Goli

Karishma Arja

COMP6115 – Spring 2015

04/28/2015

CONTENTS: **pg.no**

I. INTRODUCTION

1.1 SOCIAL NETWORKING SENTIMENT ANALYSIS

Social Media has become one of the most popular platforms to allow users discussing, communicating, and sharing their interested topics without having same geo-location and same time. Information can be generated and managed through either computer or mobile devices by one person and consumed by many other persons. Different people could express different opinions on the same topic. A wide variety of topics, ranging from current events and political debate, to sports and entertainment are being actively discussed on these social forums. The power of social media as a marketing tool has been recognized, and is being actively used by governments, major organization, schools and other groups to effectively and quickly communicate with large numbers of people. Another important metric for business to measure their online reputation is word of mouth publicity. Word of mouth is the process of spreading information from person to person, and is often done through social media networks. It also plays a major role in customer buying decisions. Collecting and analyzing these data could help users or managers make informed decisions. Marketing leaders or product managers might collect and analyze feedbacks and comments on campaigns launched by them from Facebook aiming to adopt efficient advertising strategy and improve product quality. Most of these user generated content are textual information. The rapid growth in volume of web texts from major social network sites like Facebook and Twitter drives us to analyze and mine the data through computational techniques. Identifying their sentiments has become an important issue and attracted many attentions. Recently, there have been a number of studies attempting to model/predict real-world events using information from social media networks. Among these, Twitter has attracted additional attention because of the huge surge in its popularity. Some studies conclude that 19% of tweets contain brand references, of which nearly 20% contain sentiments about the brands. The main research efforts on sentiment analysis done previously can be classified into 2 branches. On one hand, they take state-of-the-art sentiment identification algorithms to solve problems in real applications such as summarizing customer reviews, ranking products, finding product features that imply opinions. Now-a-days we analyze tweet sentiments about movies and attempts to predict box office revenue. On the other hand, researchers put their focus on discovering new sentiment algorithms. Bag-of-Words approach produces domain-specific lexicons. Some researchers propose compositional semantics, which is based on the assumption that the meaning of a compound expression is a function of the meaning of its parts and of the syntactic rules by which they are combined. They have developed a set of composition rules to assign sentiments to individual clauses, expressions and sentences. We have augmented these rules by adding our own rules which are specific to social media texts. In addition to these rules, we require a method of assessing the impact of these on the polarity of an expression. We also develop our version of the Compose function for computing the polarity of an expression based on Compositional Semantic rules.

1.2 DATA MINING

The major reason that data mining has attracted a great deal of attention in the information industry in recent years is due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. The

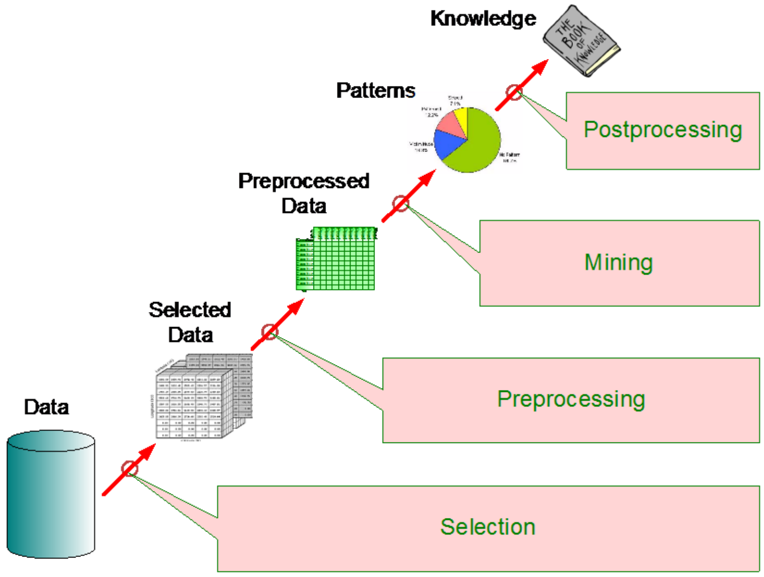


Fig 1.2 Different steps involved in Datamining.

information and knowledge gained can be used for applications ranging from business management, production control and market analysis to engineering design and science exploration.

Data mining refers to extracting or mining knowledge from large amounts of data. The output of data mining is interesting patterns in view of the user. Data must be preprocessed in order to perform any data mining functionality like classification, clustering, Association analysis etc.,. Data preprocessing includes data cleaning, data reduction, data transformation and data integration.

1.3 DATA MINING TECHNIQUES:

- Classification

Classification is the process of partitioning a given dataset into disjoint classes using a class attribute. For example, in determining a store location, the success of a store is determined by its neighborhood. The company is interested in identifying neighborhoods that would constitute its primary candidates. A model is built based on the values of all attributes to classify each item into a particular class. The goal of classification is to analyze the training set and to develop an accurate description or model for each class using the attributes presented in the data. Many classifications models have been developed such as neural networks, genetic models, and decision trees etc.

- Clustering

Clustering is the process of grouping the data into clusters with high intra-cluster similarity and low inter-cluster similarity. A similarity measure needs to be defined and the quality of the cluster, to a large extent, depends on the appropriateness of the similarity measure for the data set or the domain of application. The technique of clustering, for example, can be used to divide the market into distinct groups, so that each group can be targeted with a different strategy. There are several clustering techniques: partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model based methods.

The basic difference between classification and clustering is that classification is a supervised learning method, which assumes predefined class labels, while clustering is an unsupervised learning method that does not assume any knowledge of classes.

1.4 WEB MINING AND OPINION MINING

Web mining aims to discover useful knowledge from Web hyperlinks, page content and usage log. Based on the primary kind of data used in the mining process, Web mining tasks are categorized into three main types: Web structure mining, Web content mining and Web usage mining.

Opinion mining is a type of natural language processing for tracking the mood of the public about a particular product. Opinion mining, which is also called sentiment analysis, involves building a system to collect and examine opinions about the product made in blogs, posts, comments, reviews. Automated opinion mining often uses machine learning, a component of artificial intelligence.

1.5 MACHINE LEARNING

Machine learning, a branch of artificial intelligence, is a scientific discipline concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data, such as from_sensor data or databases. A learner can take advantage of examples (data) to capture characteristics of interest of their unknown underlying probability distribution. Data can be seen as examples that illustrate relations between observed variables. A major focus of machine learning research is to automatically learn to recognize complex patterns and make intelligent decisions based on data; the difficulty lies in the fact that the set of all possible behaviors given all possible inputs is too large to be covered by the set of observed examples (training data). Hence the learner must generalize from the given examples, so as to be able to produce a useful output in new cases.

1.6 SOCIAL NETWORKING SITES

Social Media is becoming major and popular technological platform that allows users discussing and sharing information. Information is generated and managed through either computer or mobile devices by one person and consumed by many other persons. Most of these user generated content are textual information, as Social Networks (Facebook, LinkedIn, Yelp), Microblogging (Twitter), blogs (Blogspot, Wordpress).

Looking for valuable nuggets of knowledge, such as capturing and summarizing sentiments from these huge amounts of data could help users make informed decisions. As the audience of microblogging platforms and services grows day by day, data from these sources can be used in opinion mining and sentiment analysis tasks. Microblogging platforms are used by different people to express their opinion about different topics, thus it is a valuable source of people's opinions.

For example, manufacturing companies may be interested in the following questions:

- What do people think about our product (service, company,…etc.)?
- How positive (or negative) are people about our product?
- What would people prefer our product to be like?

All this information can be obtained from microblogging services, as their users post everyday what they like/dislike, and their opinions on many aspects of their life.

1.7 Yelp

Yelp contains an enormous number of text posts and it grows every day. The collected corpus can be arbitrarily large.

We use Yelp's recently released academic dataset, which provides over one hundred fifty thousand reviews and their corresponding ratings for restaurants centered near many different universities. The data also includes other information collected by Yelp, including "useful," "funny," and "cool" ratings that is specific to each review. For the sake of training, we did not consider correspondences between the 2 users or the restaurants with reviews because we wanted to provide a purely objective analysis of semantic. Dataset main contains the following data:

1. Texts containing positive emotions, such as happiness, amusement or joy.
2. Texts containing negative emotions, such as sadness, anger or disappointment.
3. Objective texts that only state a fact or do not express any emotions.

We perform linguistic analysis of our corpus and we show how to build a sentiment classifier that uses the collected corpus as training data.

2. LITERATURE SURVEY

2.1 RELATED WORK

Recently, there has been a wide range of research done on sentiment analysis, from rule-based, bag-of-words approaches to machine learning techniques. One of the main directions is sentiment classification, which classifies the whole opinion document (e.g., a product review) as positive or negative. They also find that “content-word negators” plays an important role in determining expression-level polarity. Some authors employ machine learning techniques to classify documents by overall sentiments and conducted their experiments on movie reviews and the results show that three machine learning methods they employed (Naïve Bayes, maximum entropy classification, and support vector machines) do not perform as well on sentiment classification as on traditional topic-based categorization. Another important direction is classifying sentences as subjective or objective, and classifying subjective sentences or clauses as positive or negative. Some present a linguistic analysis of conditional sentences, and build some supervised learning models to determine if sentiments expressed on different topics in a conditional sentence are positive, negative or neutral. Several researchers also studied feature/topic-based sentiment analysis. Their objective is to extract topics or product features in sentences and determine whether the sentiments expressed on them are positive or negative. Some authors aim to summarize all customer reviews of a product by mining the features of the product on which customers have expressed their opinions and whether the opinions are positive or negative. Some authors use feature-based opinion mining model to identify noun product features that imply opinions. It is mainly focusing on the problem of objective nouns and sentences with implied opinions. Some propose an approach to extracting adverbadjective-noun phrases based on clause structure obtained by parsing sentences into a hierarchical representation. They also propose a robust general solution for modeling the contribution of adverbials and negation to the score for degree of sentiment. This is the basis of our second algorithm plus some extra phrases added by us.

2.2 EXISTING SYSTEM

In today's dynamic and global environment, social networking sites emerged like a major aspect of communication. Rating for a product will be provided by means of some special symbols like stars, thumbs up and down. This may give the rating but can't clearly exhibit the user's view and degree of satisfaction about the product. The enterprise can't even know the demand of a specific product. Similarly this applies for every practical aspect like review about a political party, movie.

Especially this can't identify the strength of the sentiment from informal text that has been used extensively in social networking sites. Existing system can't totally serve as a correct platform for accurate sentiment analysis. This can just represent the rating. It can't even represent feature selection concept which helps in detailed analysis of the product.

3. PROBLEM SPECIFICATION

The problem here is to investigate the sentiment of various products. Reviews play a prominent role in the promotion of products. Hence, we have developed algorithms based on machine learning to identify the nature of review that is either positive or negative or neutral and the degree of polarity possessed by each review.

This could allow manufacturers as well as customers to know the rating of a product as well as the demand of specific features of the product. The only channel that can enable a vast coverage of reviews is social networking sites that are the reason which made us to consider social networking sites. We have implemented the algorithms to identify strength of sentiment based on Bayes classification.

4. SOFTWARE REQUIREMENT SPECIFICATION

4.1 INTRODUCTION

A Software Requirements Specification (SRS) is a complete description of the behavior of the system to be developed. It includes a set of use cases that describe all the interactions the users will have with the software. Use cases are also known as functional requirements. In addition to use cases, the SRS also contains nonfunctional requirements. Non-functional requirements are requirements which impose constraints on the design or implementation such as performance engineering requirements, quality standards, etc.

4.1.1 Purpose:

If the organization employs this software, they can easily analyze the demand of their products. This software also enables the customers to know the reviews about desired products, Political parties may be interested to know if people support their program or not. Social organizations may ask people's opinion on current debates.

4.1.2 Scope:

The major functions of the system include Preprocessing and sentiment analysis algorithms. This enables to analyze sentiment strength from given reviews. The system reacts to the user with an output file that contains sentiment strength of given reviews. The system even notifies the user by displaying the result in the text area on the screen. The scope of this software can be extended even to general informal text other than political, movie, product reviews. Since natural language processing has been followed.

4.2 FUNCTIONAL REQUIREMENTS

It defines a function of a software system or its component. A function is described as a set of inputs, the behavior, and outputs.

- **Inputs:** User submits a file that is to be projected to sentiment analysis and selects the corresponding preprocessing task, he will be provided with the feasibility to select more than one preprocessing task.

- **Processing:** On receiving the option corresponding action is performed i.e., the software removes urls, special characters and question words from the input file. The result obtained will act as input to the algorithm. User selects the algorithm which he/she wants to implement. The software performs the sequence of steps to get the strength of sentiment in the file. It automatically performs clustering by means of which user can get separate files for positive, negative and neutral tweets. The user can also avail the facility of feature selection to identify the demand of a specific feature of a product in the market.
- **Outputs:** User can view the output file to know the sentiment strength of the reviews. He also has the facility to view the output in the text area that is displayed on the output screen.

4.3 NON FUNCTIONAL REQUIREMENTS It is a requirement that specifies the criteria that can be used to judge the operation of a system, rather than specific behaviors.

4.3.1 User Requirements

- Minimum understandability of English.
- Minimum knowledge about operating of computer.

4.3.2 Software Requirements

- Any 32 bit windows platform.
- JDK 1.6.0

4.3.3 Hardware Requirements

- Pentium IV processor.
- A minimum of 32 MB RAM.and mminimum of 2GB Hard Disk Space

5. ANALYSIS

5.1 PROPOSED SYSTEM

The proposed system has the features of preprocessing which enables accurate analysis of sentiment in the review. We maintain a database which contains words and their corresponding polarity values, preprocessed data can be efficiently searched and find the polarity. For example the word happyyyy is not in the database hence it will be preprocessed by removing noise i.e., repeated letters in the word. Now the word happy can be found in the database. In the case of urls a URL does not contribute on the sentiment of the review, hence they can be removed. Similarly questions words or 'wh' words like what, where…etc and special characters like .,/,[,],…etc will not effect on the sentiment of the review hence they are removed from the sentence.

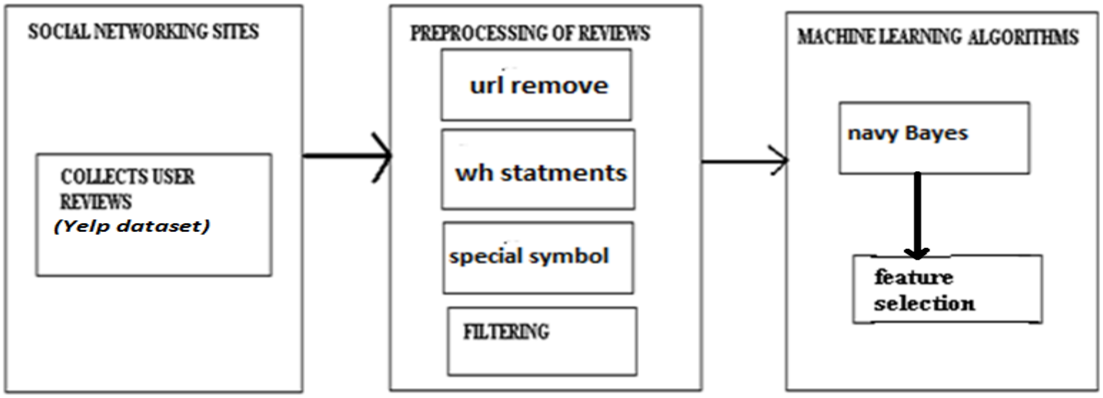


Fig 5.1 Different levels of sentiment analysis

After preprocessing the file is send as input to any one of the four algorithms that we have developed. The choice of the algorithm is left to the user. All the algorithms perform similar step initially i.e., identify each review separately and then each word separately. Find the polarity of each word from the database and calculate the polarity of each review by the techniques followed in respective algorithms. The reviews are clustered according to the polarity values in to three clusters they are positive, negative and neutral. The algorithms even let the user know number of positive reviews, negative reviews and neutral reviews in the entire file. The user can even avail the facility of feature selection by making use of one of the algorithms.

5.2 FEASIBILITY STUDY

The analysis of the existing system has to be carried to learn the details of the existing system. System analysis is the process of gathering and interpreting facts, diagnosing problems and using the information to recommended improvements to the system. Only after the systems analysis we can begin to determine how and where a computer information system can benefit all the users of the system. This system accumulation of the system called a system study.

5.2.1 Economical Feasibility Economic analysis is the most frequently used method for evaluating the effectiveness of a candidate system. More commonly known as cost/benefit analysis, the procedure is to determine the benefits and savings that are expected from a candidate system and compare them with costs. If benefits outweigh costs, then the decision is made to design and implement the system. Otherwise, further justification or alterations in the proposed system will have to be made if it is to have a chance of being approved. This is an ongoing effort that improves in accuracy at each phase of the system life cycle.

5.2.2 Technical Feasibility

Technical feasibility centers on the existing computer system (hardware, software, etc.) and to what extent it can support the proposed addition. For example, if the current computer is operating at 80 percent capacity – an arbitrary ceiling –then running another application could overload the system or require additional hardware.

This involves financial considerations to accommodate technical enhancements. If the budget is a serious constraint, then the project is judged not feasible.

5.2.3 Operational Feasibility

Purpose projects are beneficial only if they can be turned into information systems that will meet the organizations operating systems.

Some of the conditions are to check for sufficient support for the project from management and users, to check whether current business methods acceptable to the users, to check whether the users been involved in the planning and development of the project, to check whether the proposed system cause harm.

5.3 TECHNICAL ANALYSIS

People from long time have the zeal to know the sentiment of the natural language i.e., from the informal comments of a user. The enterprise can identify the user's opinions about a specific product and can even know the features. The computational treatment of opinion, sentiment, and subjectivity has recently attracted a great deal of attention (see references), in part because of its potential applications. For instance, information extraction and question-answering systems could flag statements and queries regarding opinions rather than facts (Cardie et al., 2003). Also, it has proven useful for companies, recommender systems, and editorial sites to create summaries of people's experiences and opinions that consist of subjective expressions extracted from reviews (as is commonly done in movie ads) or even just a review's *polarity* — positive ("thumbs up") or negative ("thumbs down"). Document polarity classification poses a significant challenge to data-driven methods, resisting traditional text-categorization techniques.

6. IMPLEMENTATION

6.1 SENTIMENT ANALYSIS USING SUPERVISED LEARNING

Sentiment analysis software has been built which attempt to quantify opinion from product reviews. Identifying the polarity and degree of polarity from a review has become one of the leading trends which enable the enterprises, Political parties, movie units etc., to know the opinin of the public. Hence opinion mining has become very crucial now-a-days.

This has mainly two tasks. They are preprocessing and Sentiment strength detection.

6.1.1 Preprocessing: Data preprocessing is done to eliminate the incomplete, noisy and inconsistent data. Data must be preprocessed in order to perform any data mining functionality.

Data Preprocessing involves the following tasks

- **Removing URLs**

In general URLs does not contribute to analyze the sentiment in the informal text. For example consider the sentence "I have logged in to www.Ecstasy.com as I'm bored" actually the above sentence is negative but because of the presence of the word ecstasy it may become neutral and it's a false prediction. In order to avoid this sort of failures we must employ a technique to remove URLs.

- **Filtering**

Usually people use repeated letters in words like **happyyyyy** to show their intensity of expression. But, these word are not present in the sentiwordnet hence the extra letters in the word must be eliminated. This elimination follows the rule that a letter can't repeat more than three times hence can eliminate such letter.

- **Questions**

The question words like what, which, how etc., are not going to contribute to polarity hence in order to reduce the complexity such words are removed.

- **Removing Special Characters**

Special characters like.,[]{}()/' should be removed in order to remove discrepancies during the assignment of polarity. For example "it's good:" if the special characters are not removed sometimes the special characters may concatenate with the words and make those words unavailable in the dictionary. In order to overcome this we remove special characters.

6.2 SENTIMENT ANALYSIS BASED ON SENTIWORDNET

The approach described in this paper is based on SentiWordNet, a lexical resource for opinion mining. In SentiWordNet (http://sentiwordnet.isti.cnr.it/), to each synset of WordNet, a triple of polarity scores is assigned i.e., a positivity, negativity and objectivity score. The sum of these scores is always 1. For example the triple {0, 1, 0} (positivity, negativity, objectivity) is assigned to the synset of the term "bad". The sum of all scores of this synset is 1. SentiWordNet has been created automatically by means of a combination of linguistic and statistic classifiers. It has been applied in different opinion-related tasks, i.e. for subjectivity analysis and sentiment analysis with promising results.

6.3 Bigrams:

Bigrams are used in order to increase the accuracy of the classifier. The effect of previous word on current word plays major role in sentiment analysis hence we consider bigrams rather than unigrams. In general preceding word will show more effect on the current word rather than the succeeding word hence we consider the polarity of preceding word.

For example consider the sentence "The art shows the culture and social issues prevailing at that point of time." Bigrams can be done as follows "The art", "art shows", "shows the", "culture and", "and social", "social issues", "issues prevailing" , "prevailing at", and so on.

6.4 DATASET

The dataset consists of series of reviews about any product. The example dataset is shown below about a restaurant in YELP. (reviews after preprocessed)

*I really love the **family style** of sharing and food coming out when it is ready instead of staged courses.*

*I really liked the brussels sprouts and beef bavette but another clear winner was the **duck breast**.*

***Chef Gary** is a master of his trade, which is fresh Cajun cooking.*

*I had the Bayou Rolls and they were awesome with a side of collard greens and **grilled cabbage**.*

*From our awesome waitress to eating phenomenal **alligator stew**, this was a highlight of our vacation.*

*The **Hanid and Rice** made me realize how I've missed goat meat.*

***Muhammad** is a very nice guy and is passionate about his food.*

*The **Fuul** is a dip/soup made with fava beans served with wheat french bread*

*Beans and slaw were fantastic, steak fries are **steak fries**.*

6.4.1 YELP DATASET Format

```
{
 'type': 'review',
  'business_id': (the identifier of the reviewed business),
  'user_id': (the identifier of the authoring user),
  'stars': (star rating, integer 1-5),
  'text': (review text),
  'date': (date, formatted like '2011-04-19'),
  'votes': {
    'useful': (count of useful votes),
    'funny': (count of funny votes),
    'cool': (count of cool votes)
  }
}
```

6.5 DETECTION OF SENTIMENT STRENGTH FROM INFORMAL TEXT IN MACHINE LEARNING APPROACH (ALGORITHM):

A Naive Bayes classifier is a probabilistic model based on Bayes' theorem. Bayes' theorem is formulated as follows.

P(c/d)=(P(c)P(d/c))/P(d)

Consider the problem of classifying documents by their content, for example into positive and negative reviews. Imagine that document are drawn from a number of classes of documents which can be modeled as sets of words where the(independent) probability that i_th word of a given document occurs in a document from class c can be written as

$$p(w_i|C)$$

(for this treatment, we simplify things further by assuming that words are randomly distributed in the document-that is, words are not dependent on the length of the document, position within the document, with relation to other words, or other document-context)

Then the probability of given document D contains all the words Wi , given a class C is

$$p(D|C) = \prod_i p(w_i|C)$$

The question that we desire to answer is:"what is the probability that a given document D belongs to a given class C ?"

In other words, what is $p(C|D)$?

Now by definition

$$p(D|C) = \frac{p(D \cap C)}{p(C)}$$ And

$$p(C|D) = \frac{p(D \cap C)}{p(D)}$$

Bayes theorem manipulates these into a statement of probability in terms of likelihood.

$$p(C|D) = \frac{p(C)}{p(D)}\, p(D|C)$$

Assume for the moment that these are only two mutually exclusive classes S and ~S (eg. Positive and negative), such that every element (review) is in either one or the other;

$$p(D|S) = \prod_i p(w_i|S)$$

And

$$p(D|\neg S) = \prod_i p(w_i|\neg S)$$

Using the Bayesian result above, we can write

Dividing one by the other gives:

$$\frac{p(S|D)}{p(\neg S|D)} = \frac{p(S)\ \prod_i p(w_i|S)}{p(\neg S)\ \prod_i p(w_i|\neg S)}$$

Which can be refactored as :

$$\frac{p(S|D)}{p(\neg S|D)} = \frac{p(S)}{p(\neg S)} \prod_i \frac{p(w_i|S)}{p(w_i|\neg S)}$$

Thus the probability ratio P(S/D)/P(~S/D) can be expressed in terms of a series of likeli-hood ratios.The actual probability P(S/D) can be easily computed for log(P(S/D)/P(~S/D)) based on the observation that P(S/D)+P(~S/D)=1

Taking the logarithm on all these ratios, we have:

$$\ln \frac{p(S|D)}{p(\neg S|D)} = \ln \frac{p(S)}{p(\neg S)} + \sum_i \ln \frac{p(w_i|S)}{p(w_i|\neg S)}$$

$$p(S|D) = \frac{p(S)}{p(D)} \prod_i p(w_i|S)$$

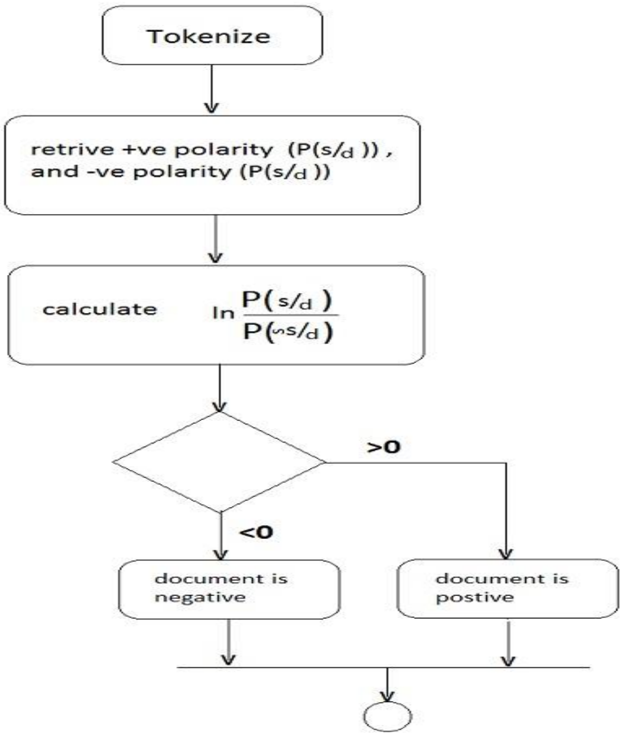
$$p(\neg S|D) = \frac{p(\neg S)}{p(D)} \prod_i p(w_i|\neg S)$$

Finally, the document can be classified as follows,

It is positive if P(S/D)>P(~S/D) (ie., $\ln \frac{p(S|D)}{p(\neg S|D)} > 0$),

It is negative if P(S/D)<P(~S/D),

otherwise, it is neutral.


Tokenize
retrive +ve polarity (P(s/d)) , and -ve polarity (P(s/d))
calculate $\ln \frac{P(s/d)}{P(\sim s/d)}$
>0
<0
document is negative
document is postive

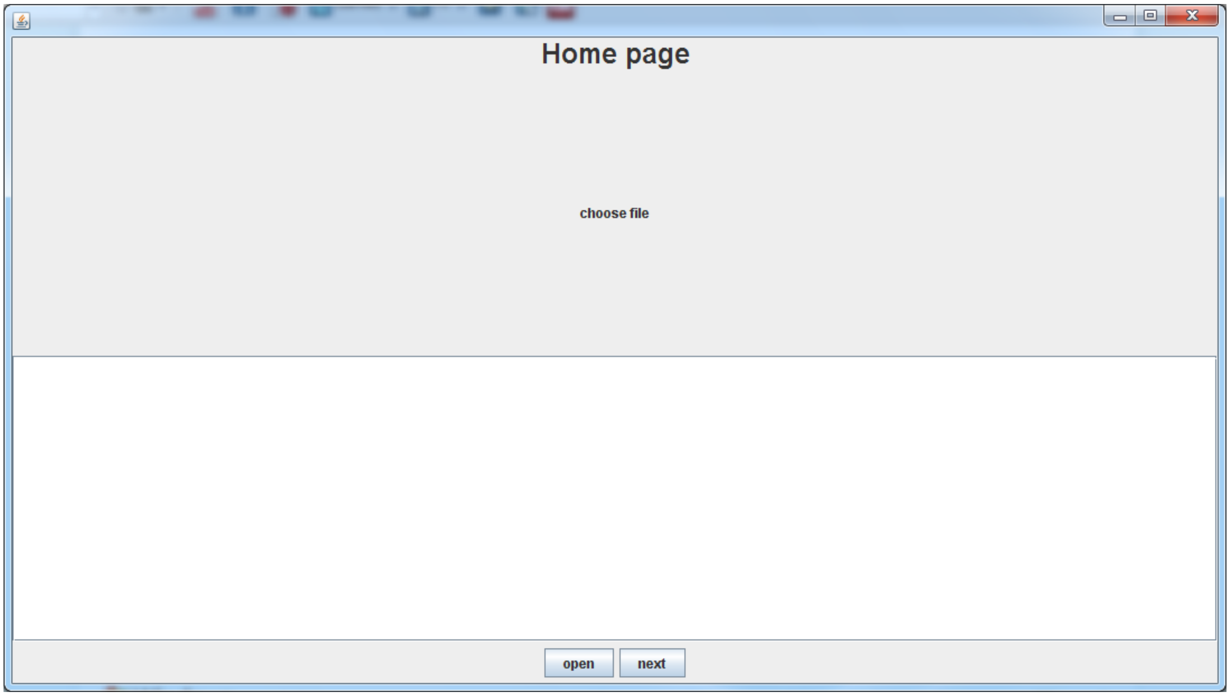
6.5 flowchart for navie Bayes

6.6FEATURE SELECTION

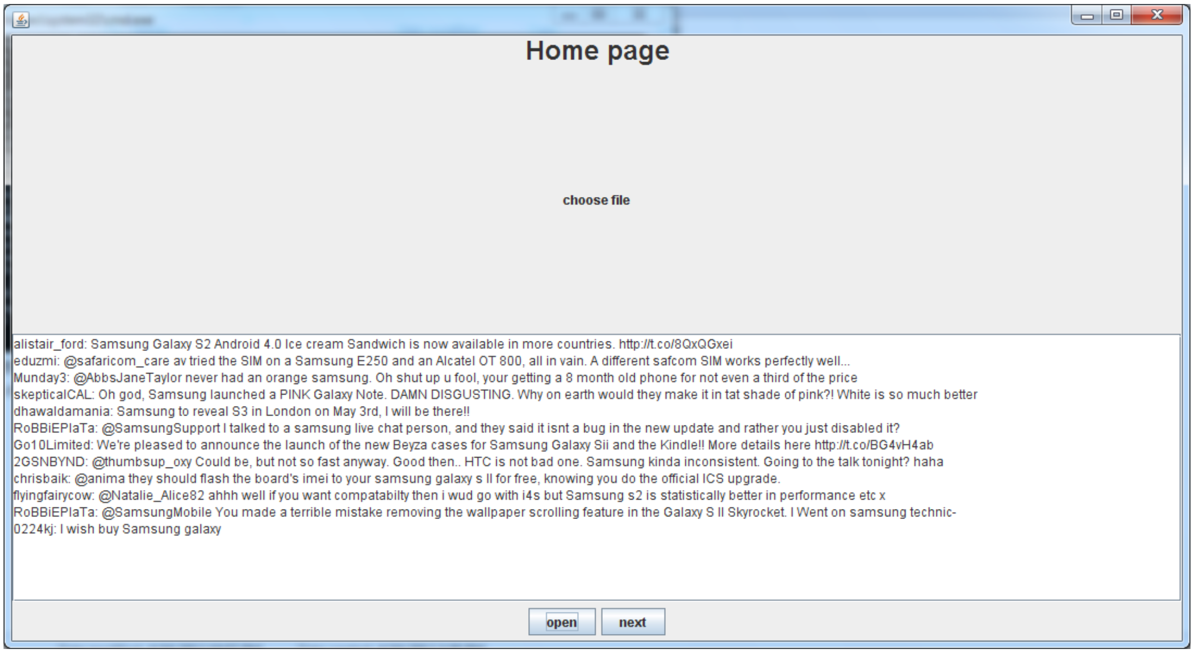
This became one of the revolutionary aspect in the present day industrial development as well as in various social scenarios like politics, movies etc.,. By using this algorithm we can identify the feature of the product which is under demand and which is under negative shade. So that the enterprise can improvise that specific feature.

- The major features of a product are tabulated in a database.
- Each and every word in the review will be compared to the features in the database.
- If the word is found then the count for that corresponding feature is incremented.
- This count symbolizes the demand for a specific feature.
- Nature of the feature can be identified by giving the output file of clustering as input to feature selection.
- If the count of that feature in negative cluster is more then it will be concluded that the feature is a failure otherwise if the count of that feature in positive cluster is more, then the feature is considered as success.

7. OUTPUT SCREENS



7.1 home page



```
alistair_ford: Samsung Galaxy S2 Android 4.0 Ice cream Sandwich is now available in more countries. http://t.co/8QxQGxei
eduzmi: @safaricom_care av tried the SIM on a Samsung E250 and an Alcatel OT 800, all in vain. A different safcom SIM works perfectly well...
Munday3: @AbbsJaneTaylor never had an orange samsung. Oh shut up u fool, your getting a 8 month old phone for not even a third of the price
skepticalCAL: Oh god, Samsung launched a PINK Galaxy Note. DAMN DISGUSTING. Why on earth would they make it in tat shade of pink?! White is so much better
dhawaldamania: Samsung to reveal S3 in London on May 3rd, I will be there!!
RoBBiEPlaTa: @SamsungSupport I talked to a samsung live chat person, and they said it isnt a bug in the new update and rather you just disabled it?
Go10Limited: We're pleased to announce the launch of the new Beyza cases for Samsung Galaxy Sii and the Kindle!! More details here http://t.co/BG4vH4ab
2GSNBYND: @thumbsup_oxy Could be, but not so fast anyway. Good then.. HTC is not bad one. Samsung kinda inconsistent. Going to the talk tonight? haha
chrisbaik: @anima they should flash the board's imei to your samsung galaxy s II for free, knowing you do the official ICS upgrade.
flyingfairycow: @Natalie_Alice82 ahhh well if you want compatabilty then i wud go with i4s but Samsung s2 is statistically better in performance etc x
RoBBiEPlaTa: @SamsungMobile You made a terrible mistake removing the wallpaper scrolling feature in the Galaxy S II Skyrocket. I Went on samsung technic-
0224kj: I wish buy Samsung galaxy
```

7.2 Loading file

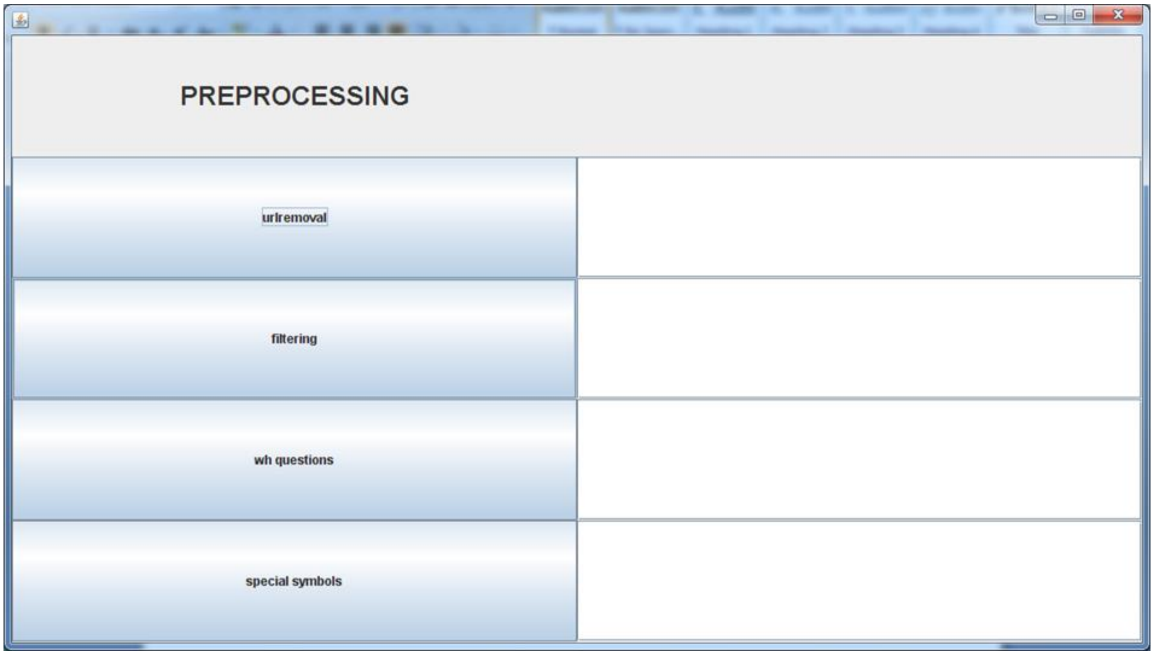
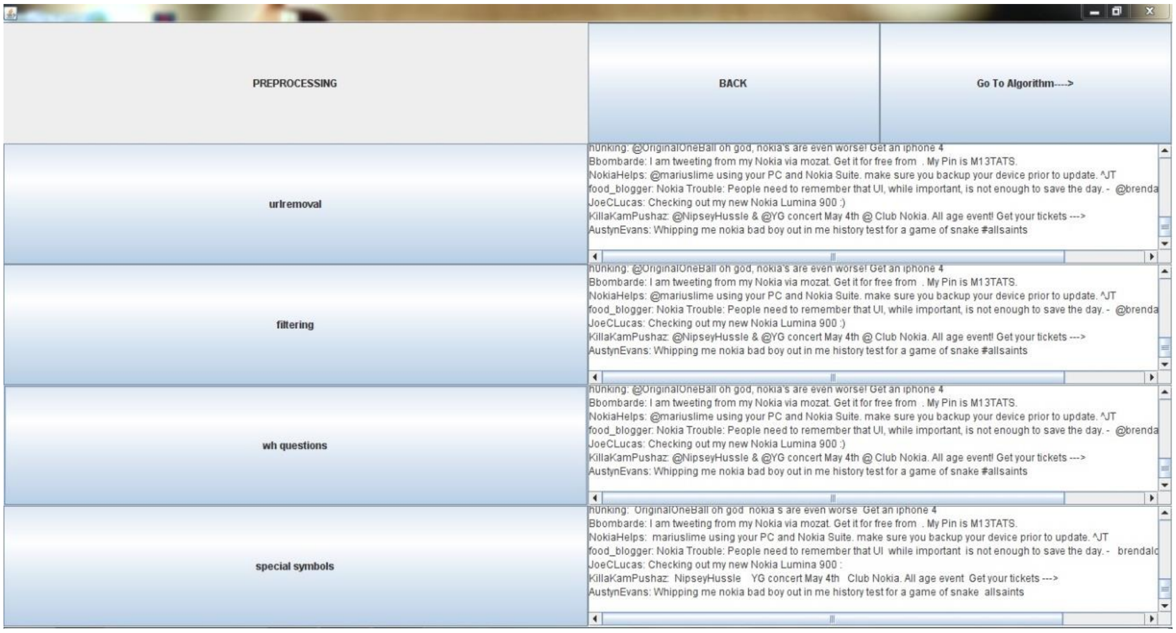


Fig 7.3 Preprocessing window with options to select preprocessing task



```
h0nking: @OriginalOneBall oh god, nokia's are even worse! Get an iphone 4
Bbombarde: I am tweeting from my Nokia via mozat. Get it for free from  . My Pin is M13TATS.
NokiaHelps: @mariuslime using your PC and Nokia Suite. make sure you backup your device prior to update. ^JT
food_blogger: Nokia Trouble: People need to remember that UI, while important, is not enough to save the day. -  @brenda
JoeCLucas: Checking out my new Nokia Lumina 900 :)
KillaKamPushaz: @NipseyHussle & @YG concert May 4th @ Club Nokia. All age event! Get your tickets --->
AustynEvans: Whipping me nokia bad boy out in me history test for a game of snake #allsaints
```

```
h0nking: @OriginalOneBall oh god, nokia's are even worse! Get an iphone 4
Bbombarde: I am tweeting from my Nokia via mozat. Get it for free from  . My Pin is M13TATS.
NokiaHelps: @mariuslime using your PC and Nokia Suite. make sure you backup your device prior to update. ^JT
food_blogger: Nokia Trouble: People need to remember that UI, while important, is not enough to save the day. -  @brenda
JoeCLucas: Checking out my new Nokia Lumina 900 :)
KillaKamPushaz: @NipseyHussle & @YG concert May 4th @ Club Nokia. All age event! Get your tickets --->
AustynEvans: Whipping me nokia bad boy out in me history test for a game of snake #allsaints
```

```
h0nking: @OriginalOneBall oh god, nokia's are even worse! Get an iphone 4
Bbombarde: I am tweeting from my Nokia via mozat. Get it for free from  . My Pin is M13TATS.
NokiaHelps: @mariuslime using your PC and Nokia Suite. make sure you backup your device prior to update. ^JT
food_blogger: Nokia Trouble: People need to remember that UI, while important, is not enough to save the day. -  @brenda
JoeCLucas: Checking out my new Nokia Lumina 900 :)
KillaKamPushaz: @NipseyHussle & @YG concert May 4th @ Club Nokia. All age event! Get your tickets --->
AustynEvans: Whipping me nokia bad boy out in me history test for a game of snake #allsaints
```

```
h0nking:  OriginalOneBall oh god  nokia s are even worse  Get an iphone 4
Bbombarde: I am tweeting from my Nokia via mozat. Get it for free from  . My Pin is M13TATS.
NokiaHelps:  mariuslime using your PC and Nokia Suite. make sure you backup your device prior to update. ^JT
food_blogger: Nokia Trouble: People need to remember that UI  while important  is not enough to save the day. -   brendald
JoeCLucas: Checking out my new Nokia Lumina 900 :
KillaKamPushaz:  NipseyHussle   YG concert May 4th   Club Nokia. All age event  Get your tickets --->
AustynEvans: Whipping me nokia bad boy out in me history test for a game of snake  allsaints
```

Fig 7.4 Preprocessing window that displays results of corresponding tasks

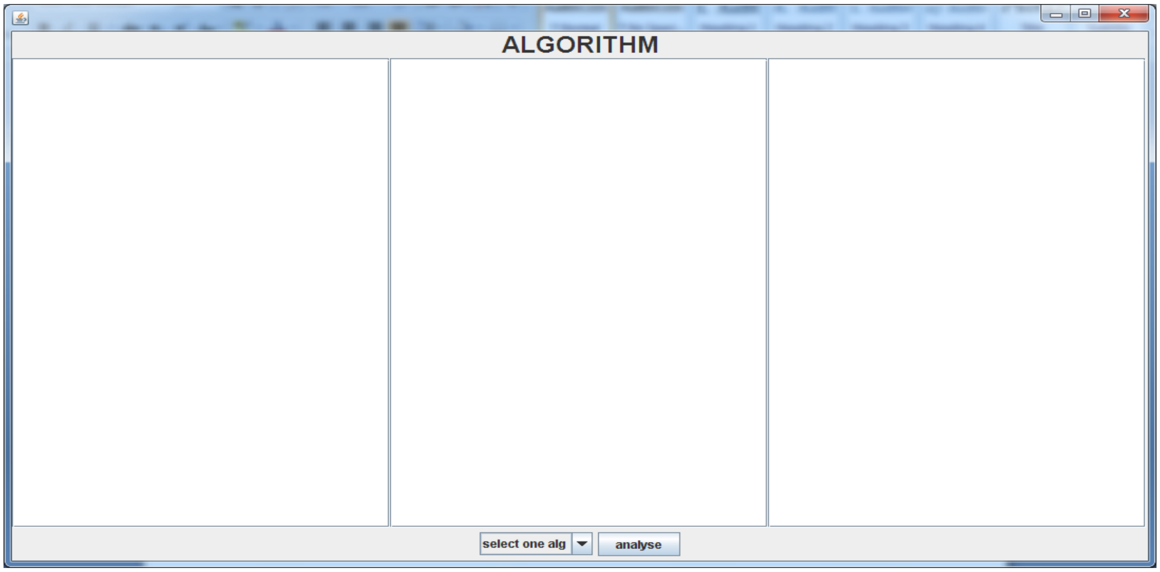


Fig 7.5 Algorithm Screen

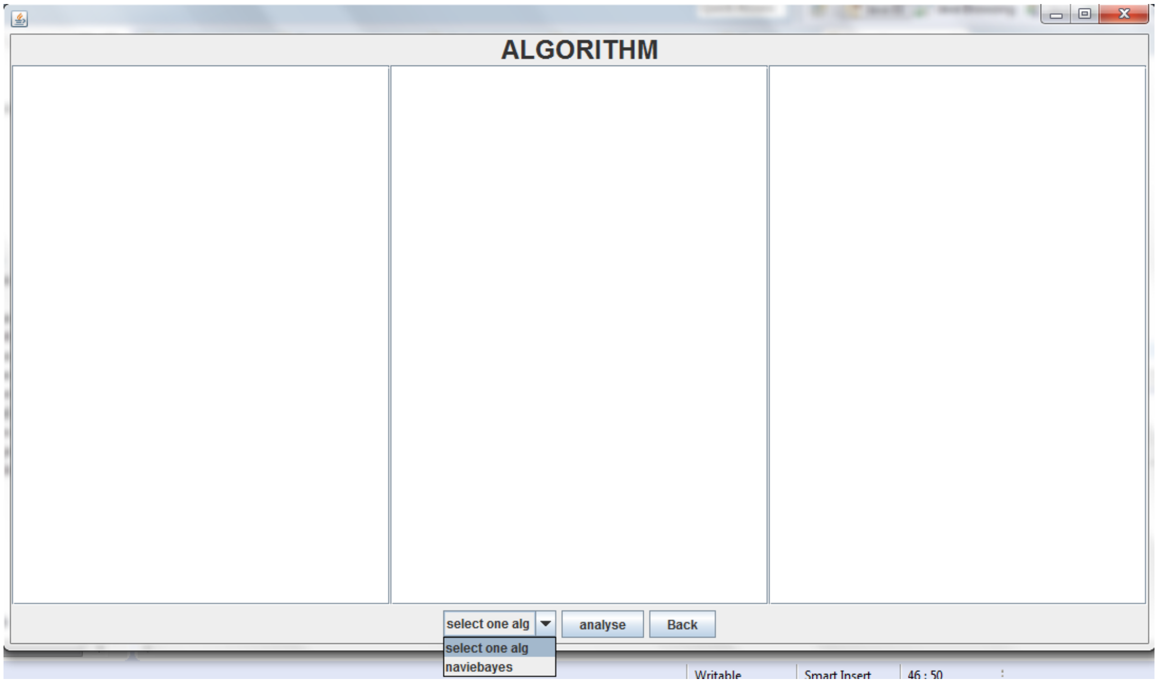
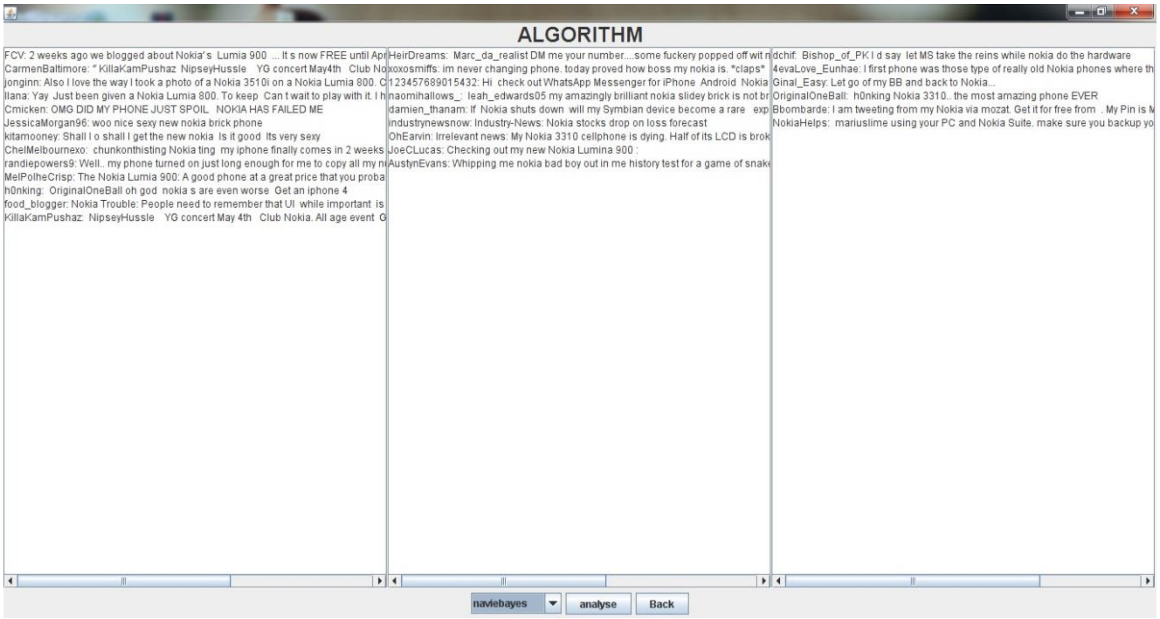


Fig 7.6 Screen that enable user to select one among various algorithm options.



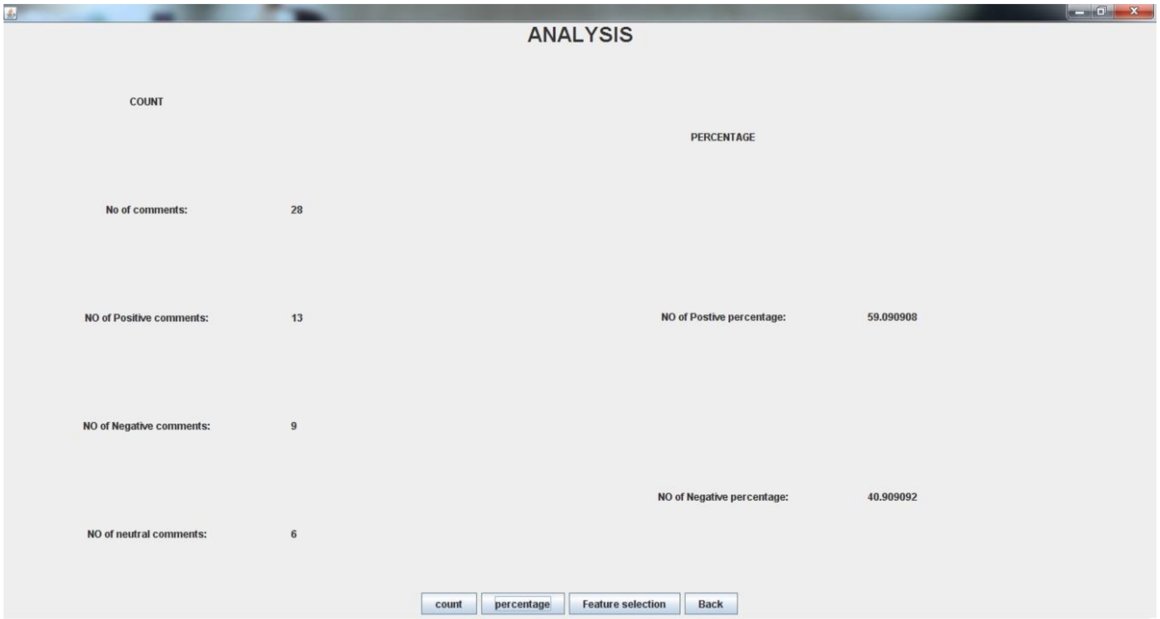
ALGORITHM

FCV: 2 weeks ago we blogged about Nokia's Lumia 900 ... It s now FREE until Apr
CarmenBaltimore: " KillaKamPushaz NipseyHussle YG concert May4th Club No
jonginn: Also I love the way I took a photo of a Nokia 3510i on a Nokia Lumia 800. C
llana: Yay Just been given a Nokia Lumia 800. To keep Can t wait to play with it. I h
Cmicken: OMG DID MY PHONE JUST SPOIL NOKIA HAS FAILED ME
JessicaMorgan96: woo nice sexy new nokia brick phone
kitamooney: Shall I o shall I get the new nokia Is it good Its very sexy
ChelMelbournexo: chunkonthisting Nokia ting my iphone finally comes in 2 weeks
randiepowers9: Well.. my phone turned on just long enough for me to copy all my n
MelPolheCrisp: The Nokia Lumia 900: A good phone at a great price that you proba
h0nking: OriginalOneBall oh god nokia s are even worse Get an iphone 4
food_blogger: Nokia Trouble: People need to remember that UI while important is
KillaKamPushaz: NipseyHussle YG concert May 4th Club Nokia. All age event G

HeirDreams: Marc_da_realist DM me your number....some fuckery popped off wit n
xoxosmiffs: im never changing phone. today proved how boss my nokia is. *claps*
12345768901 5432: Hi check out WhatsApp Messenger for iPhone Android Nokia
naomihallows_: leah_edwards05 my amazingly brilliant nokia slidey brick is not br
damien_thanam: If Nokia shuts down will my Symbian device become a rare exp
industrynewsnow: Industry-News: Nokia stocks drop on loss forecast
OhEarvin: Irrelevant news: My Nokia 3310 cellphone is dying. Half of its LCD is brok
JoeCLucas: Checking out my new Nokia Lumina 900 :
AustynEvans: Whipping me nokia bad boy out in me history test for a game of snake

dchif: Bishop_of_PK I d say let MS take the reins while nokia do the hardware
4evaLove_Eunhae: I first phone was those type of really old Nokia phones where th
Ginal_Easy: Let go of my BB and back to Nokia...
OriginalOneBall: h0nking Nokia 3310.. the most amazing phone EVER
Bbombarde: I am tweeting from my Nokia via mozat. Get it for free from . My Pin is M
NokiaHelps: mariuslime using your PC and Nokia Suite. make sure you backup yo

Fig 7.7 Screen that shows the classification of the document into positive, negative, neutral classes



ANALYSIS

COUNT

PERCENTAGE

No of comments: 28

NO of Positive comments: 13

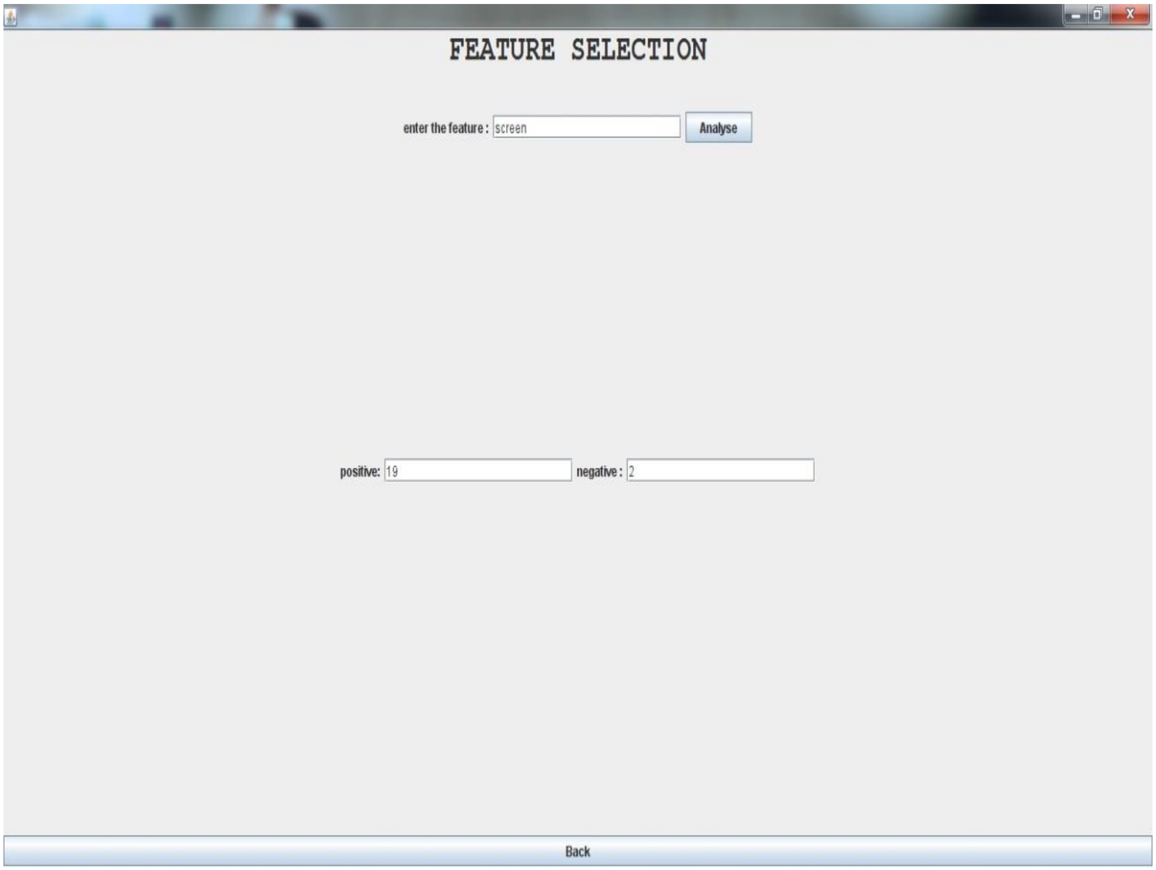
NO of Postive percentage: 59.090908

NO of Negative comments: 9

NO of Negative percentage: 40.909092

NO of neutral comments: 6

Fig 7.8 Screen that shows the analysis of the document with count and percentage values



FEATURE SELECTION

enter the feature : screen

positive: 19 negative : 2

Fig 7.9 Screen that shows the feature selection

8. USER MANUAL

- The sender first needs to open the software.
- He needs to specify the file name on which sentiment analysis is to be performed.
- He needs to specify the preprocessing tasks that are to be executed on the file.
- He clicks submit button to redirect to sentiment analysis window.
- He selects the algorithm by means of which sentiment analysis is to be performed.
- Selection of algorithm depends on the following criterion
    - Whether we need polarity or feature selection.
    - If we need to perform feature selection it is always desirable to calculate polarity and send the output file of polarity as input to feature selection algorithm.
- He gets the result in the text area as well as in the form a file.

9. CONCLUSION

One’s opinion about a topic can’t be predicted just by his/her words even expression plays a major role. Till now psychologists only have the chance to identify the opinion of an individual, with the emerging research work on sentiment analysis even a normal person can perform opinion mining.

Many organizations who want to study the pulse of public regarding their product can be known by using sentiment strength algorithms.

Before applying any classification algorithm data must be preprocessed. We have performed three preprocessing tasks. One task to remove URLs from the input file next one to remove special characters, here we can also remove repeated letters from a word, the last task is to remove question words. Now the preprocessed document can be given as input to algorithms.

We have proposed a total of four algorithms in order to find sentimental strength. Three algorithms determine the polarity of the sentences both degree as well as type of polarity i.e., either positive or negative or neutral. First algorithm works as a simulation of naïve bayes theorem i.e., it works on the basis of probability values of each and every word in the sentence. Second algorithm considers the words before and after that specified word which we are working out now. This algorithm adds the polarities of predecessor and current word, the polarity of current word and successor are added finally both the sums are compared and the greater value is given as the polarity of current word. Third algorithm includes the features of negators and boosters as well. We consider three words on both sides of the current word to get the polarity of the current word. Fourth algorithm determines the feature of the product that is under positive or negative shade. Depending on the option of negative or positive rating of features, input will be given to feature selection.

10.FUTURE ENCHANCEMENTS

The number of internet users has been increasing day by day. Social networking sites are emerging with rocket speed. The zeal to understand the opinion of an individual based on the review he had given increased. The ongoing developments in the field of opinion mining added fuel to this interest.

This is getting implemented in online reviews at present. This can be extended to short text messages that are transferred in cell phones. This enables most of the users to make use of this. This facility can be incorporated in our project by providing compatibility with the mobile technologies. This can even be used by government defense systems to detect the seriousness of the situation by tracking the conversations of unofficial organizations.

11. REFERENCES

[1]. Document-Word Co-Regularization for Semi-supervised Sentiment Analysis by Vikas Sindhwani and Prem Melville, Business Analytics and Mathematical Sciences, IBM T.J. Watson Research Center, Yorktown Heights, NY 10598 {vsindhw,pmelvil}@us.ibm.com

[2]. Fully Automatic Lexicon Expansion for Domain-oriented Sentiment Analysis by Hiroshi Kanayama Tetsuya Nasukawa, Tokyo Research Laboratory, IBM Japan, Ltd. 1623-14 Shimotsuruma, Yamato-shi, Kanagawa-ken, 242-8502 Japan {hkana,nasukawa}@jp.ibm.com

[3]. LargeScale Sentiment Analysis for News and Blogs Namrata Godbole? Manjunath Srinivasaiah? Steven Skiena_namratagodbole@gmail.com manj.blr@gmail.com skiena@cs.sunysb.edu?Google Inc., New York NY, USA}Dept. of Computer Science, Stony Brook University, Stony Brook, NY 11794-4400, USA

[4]. Paroubek, Universit´e de Paris-Sud, Laboratoire LIMSI- CNRS, Bˆatiment 508,F Twitter as a Corpus for Sentiment Analysis and Opinion Mining Alexander Pak, Patrick -91405 Orsay Cedex, France,alexpak@limsi.fr, pap@limsi.fr

[5]. Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language processing Techniques - ‡ IBM Tokyo Research Lab, 1623-14, Shimotsuruma Yamato-shi, Kanagawa-ken 242-8502, Japan nasukawa@ip.ibm.com

[6]. Sentiment Elicitation System for Social Media Data - Kunpeng Zhang, Yu Cheng, Yusheng Xie, Daniel Honbo Ankit Agrawal, Diana Palsetia, Kathy Lee, Wei-keng Liao, and Alok Choudhary - Department of Electric Engineering & Computer Science, Northwestern University, Evanston, IL 60208,

[7]. Sentiment Analysis in Practice Yongzheng (Tiger) Zhang , Dan Shen, Catherine Baudin*

[8]. Sentiment Analysisin Short and Informal Text – Marco Veluscek with the supervision of Prof. Sune Lehmann, PhD

[9]. Text normalization in social media: progress, problems and applications for a pre-processing system of casual English - Eleanor Clarka and Kenji Arakia*

[10]. Pre-processing very noisy text - Alexander Clark, ISSCO / TIM, University of Geneva, UNI-MAIL, Boulevard du Pont-d'Arve, CH-1211 Geneva 4, Switzerland, Alex.Clark@issco.unige.ch

[11]. Feature Selection and Weighting Methods in Sentiment Analysis, Tim O'Keefe, School of Information Technologies, University of Sydney, NSW 2006, Australia, toke9145@uni.sydney.edu.au, irena@it.usyd.edu.au