NEWS FILTERING VIA NAIVE BAYES TEXT CLASSIFICATION: A DATA MINING APPROACH

A Project paper submitted to
the Faculty of Computer Science and Engineering of
Hajee Mohammad Danesh Science and Technology University
in partial fulfillment of the requirements for the degree of
Bachelor of Science in Computer Science and Engineering

Submitted By
Sanjoy Roy Duke
Student ID: 0702013
Session: 2007

Supervised By
Ashis Kumar Mandal
Assistant Professor, Department of CEN

Submitted To
Department of Computer Science and Information Technology
HAJEE MOHAMMAD DANESH SCEIENCE AND TECHNOLOGY UNIVERSITY

January, 2012

NEWS FILTERING VIA NAIVE BAYES TEXT CLASSIFICATION: A DATA MINING APPROACH

DEDICATED TO -

EVERYONE WHO TRY TO NEVER COME IN FRONT OF THE PEOPLE.

IF THE PERSON BE PRESENT IN A CEREMONEY, NEVER BEEN TOLD TO SIT IN THE FIRST ROW.

IF THE PERSON BE ABSENT, NOBODY ASKS FOR.

Faculty of Computer Science and Engineering
Hajee Mohammad Danesh Science and Technology University
Dinajpur-5200, Bangladesh.

CERTIFICATE

This is to certify that, the project paper entitled **"NEWS FILTERING VIA NAÏVE BAYES TEXT CLASSIFICATION:A DATA MINING APPROACH** submitted by **SANJOY ROY DUKE**, Student ID: **0702013**, has been carried out under the Department of **Computer Science and Information Technology (CIT)** and under the supervision of **ASHISH KUMAR MANDAL, Assistant Professor, Department of Computer Engineering (CEN)**. This project and thesis has been prepared under the Department of **Computer Science and Information Technology** in **Hajee Mohammad Danesh Science and Technology University (HSTU)** as a 4.5 Credit Hour course (CIT-402, Project and Thesis) in partial fulfillment of the requirement for the Degree of **Bachelor of Science in Computer Science and Engineering**.

09/01/12

Supervisor
Ashis Kumar Mandal
Assistant Professor and Chairman
Department of CEN
HSTU, Dinajpur.

Chairman
Md. Delowar Hossain
Assistant Professor and Chairman
Department of CIT
HSTU, Dinajpur.

ACKNOWLEDGEMENT

All the praises, gratitude and thanks to GOD who enables me to purpose my education in Computer Science and Engineering (CSE) faculty and to complete this Project and Thesis paper for the degree of Bachelor of Science in CSE.

I express my heartiest gratitude and profound application to Ashis Kumar Mandal, Assistant Professor and Chairman of Computer Engineering Department of Hajee Mohammad Danesh Science and Technology University, Dinajpur-5200 for his managerial guidance, meticulous supervision, advice, providing importance literature regarding this paper, encouragement throughout the period of the study and made helpful comment on the draft through which it was possible to prepare this paper.

I divulge my gratefulness and indebtedness to all of my honorable teachers of the Faculty of CSE, Hajee Mohammad Danesh Science and Technology University, Dinajpur for their encouragement, valuable suggestions during my study period, especially during the preparation of the Project and Thesis paper.

I express heartiest gratitude and profound love to all my friends, classmates, senior brothers and sisters, junior students of our faculty for their encouragement and concern.

I also would like to thank the entire Stuff of CSE who were very helpful and cordial with me through the years and especially for successful completion of the Project.

ABSTRACT

In recent world information increased all the time. People get data from various sources. Thus the electronic media get news and provide those news to the people all the time. They have to collect information and gather data in a particular area and also need to classify that information. This process is lengthy and difficult to maintain the accuracy of the classification in a sort of time.

News providers need to handle this problem. They have to read the news deep and find the pattern quickly and correctly. But, a automated system can solve this problem in a sort time and accuracy.

News are different kinds of like text, audio, video. Most of news are text news of data that need to classify. These data are in a unstructured way, so a news filtering technique is developed to solve that problem. This is actually a text classification techniques that has been used.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1: Introduction

1.2 Overview

Of late the volume of information increases day by day. All these data are not in proper structure. This type of data are called unstructured data or information. Unstructured data is what we find in mails, messages, reports, presentations, newspapers, voice mails. These unstructured data are causing problem in recent activity.

Managing unstructured data is a problem. The huge data are available on internet, on the web and causing data overloading problem. According to Robb, "We are drawing in information but are starving for knowledge". Data are useful when it is in an organized way. For this reason managing the unstructured data is important. The unstructured data usually in a free format and mostly in the text format which is difficult to manage. From many times ago there have been so many research to solve this type of problem.

Unstructured data usually used in many circumstances, like corporate portal, knowledge management system, electronic media. In recent world electronic media deals with a huge amount of information or data. But to serve news, media must need to handle these huge unstructured data. The main problem is to handle these data and also need to find meaning from these data. So, text or news classification is important to solve this kind of problem.

Most of the text categorization are based on machine learning and statistics. According to Mitchell (1997), machine learning is a computer programs that learn from experience (E) with respect to some class of tasks (T) and performance measure (P), if it's performance at tasks in T, as measured by P, and improves with experience E. The explanations of three elements in the machine learning are as follows :

- Tasks (T): It is the kind of activity on which the computer will learn to improve its performance.
- Performance (P): A measure of the quality of the response or action.
- Experience (E): It means that what has been recorded in the past

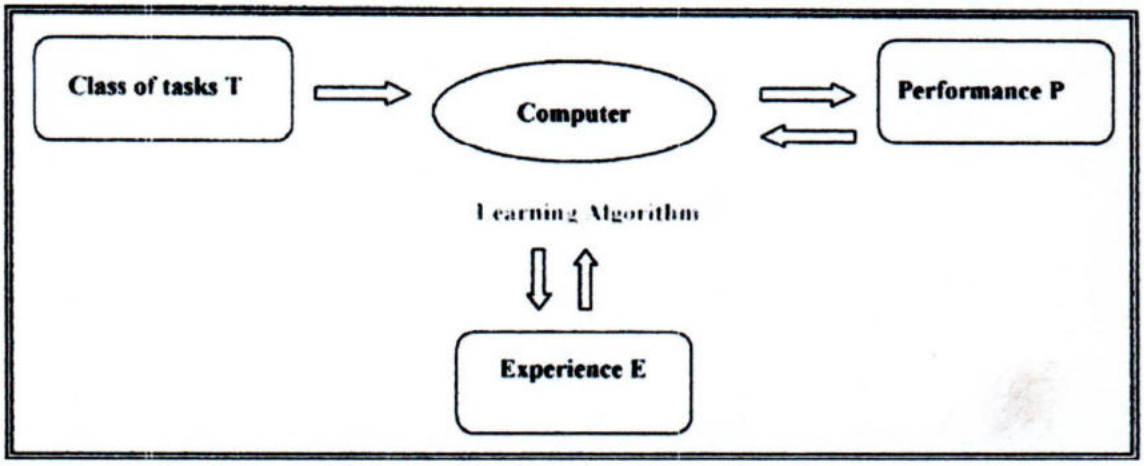

Fig.1.1: Definition of Machine Learning (Source: Mitchell, 1997)

1.2.1 NAIVE BAYES

Naïve Bayes is classification technique based on the "Bayesian probability". It is often called as naïve bayes classifier. Bayesian probability was been proposed by Thomas Bayes who proved a special case which in Bayes's theorem. This theorem first introduced in 1950. It's a simply probability classifier. It computes the likelihood that a program is malicious given the features that are contained in the program. This classifier is also a simple algorithm for categorization process and gives a good performance in getting a result (Shen & Jiang 2003). The advantages of this technique is that it's a simple but high accuracy in getting result.

1.2.2 SUPERVISED LEARNING

Supervised learning is the machine learning process of inferring a function from *supervised* (labeled) training data. The training data consist of a set of *training examples*. In supervised learning, each example is a *pair* consisting of an input object (typically a vector) and a desired output value (also called the *supervisory signal*). A supervised learning algorithm analyzes the training data and produces an inferred function, which is called a *classifier* (if the output is discrete, see classification)or a *regression function* (if the output is continuous, see regression). The inferred function should predict the correct output value for any valid input object. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way.

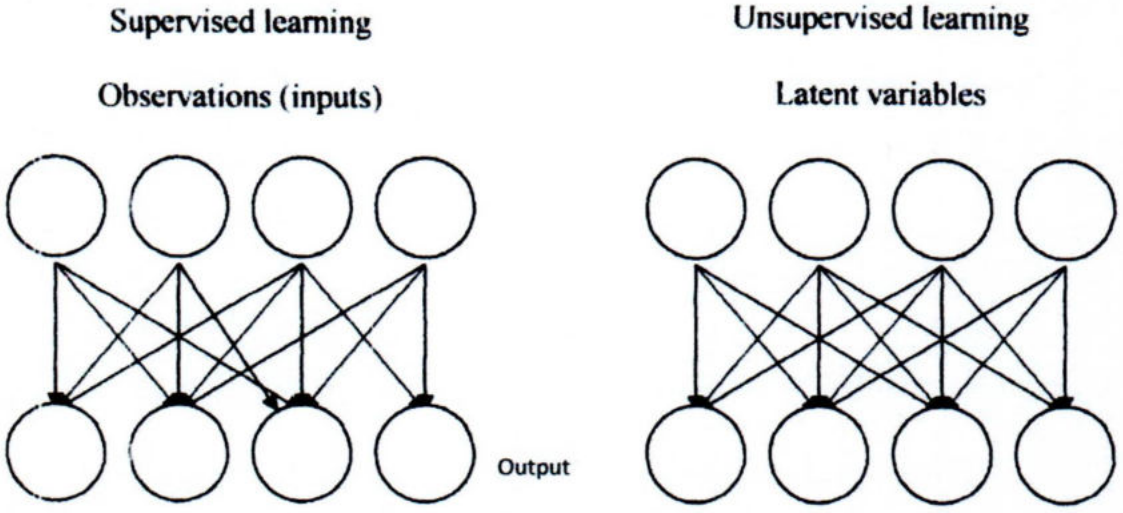

Fig.1.2: Supervised and Unsupervised Learning (Source : Valpola,2000)

1.2 Problem Statement

In electronic media the news must be in a structured way. But mainly different types of news must read and need to take a decision that what type of news is. So, need a automated system to classify news. Otherwise it will take a long time to classify the news. All these documents must check seriously so, to do this by men it's quite difficult to sort the all news in a right categories.

To sort news by men and to do this manually it will kill time and accuracy. It's also difficult to find pattern in a unstructured data. To find pattern there will be a lot of calculation will be needed. So, no one can do it with correctness. For the help of many electronic devices in a news organization there comes a lot of information in an hour. So, classification is also frequently needed.

1.3 OBJECTIVES OF STUDY

The objective of this study are as follows :

- To classify the textual document using naïve Bayes Classifier.
- To save the time and maintain the correctness of the classifier.
- Classify the news in faster way.

1.4 SCOPE OF STUDY

The scope of this study is below :

- This study focused on filtering unstructured news data.
- Focus on the text type data only.
- The text document contains unstructured data and and news for various fields.
- There must be a training part for the supervised learning.
- There must be new instances of data actually that are different types of news text data and must classify that data.
- The classification technique must follow the Naïve Bayes Text Classificaion techniques.

Chapter 2: LITERATURE REVIEW

2.1 UNSTRUCTURED DATA

Usually in text form the data are in free and unstructured form. Such form of data are in a news text data. Normally these type of data are unstructured and difficult to manage and also time consuming to retrieve a particular news. Unstructured data or news are huge compared to structured data and it's also known as bulk data. Structured data is information that has been organizer to allow and separation of the context of the information from its content. It is uniform, consistent pattern for arranging data in a file. When data is structured in a file, the data can easily used with less chance for errors. The unstructured data is more valuable data compared to structured data for the news organization. This is the cause large data can be obtained from the documents, emails, images, and other type of unstructured data.

According to Rao(2003), "the term unstructured data refers to the difficult of applying Information Technology (IT) to routing and accessing content which is slicing and dicing and manipulating it in all the ways typical of data stored in rows and columns in relation database". The unstructured data can be managed effectively and efficient by extracting the information from the data and the extracted data will be processed into structured form.

The process of extraction is the process of extracting the information from the sets of data or documents. This process allows for scanning the patterns among the words and phrases.The benefit of extraction process is people can quickly find information from the textual documents of news without reading for deeper meaning. There are two levels of extraction which are entity extraction and fact extraction. Entity extraction is focuses on identifying the entity named such as people, organizations, products, and places whereas fact extraction is the process of spreading out the facts from the entities and topics, then connecting and contextualizing the fact in relationship (Rao).

The extraction techniques can also link and contextualize topics in various ways. There is an approach for extracting information from unstructured data which is keyword frequency. This approach is used to analyze and access sets of documents using frequencies of occurance of various keywords representing the documents(Gurusamy)

2.2 NEWS CLASSIFICATION

The approaches that applied in this project are categorization. According to Rao a classification system can accurately assign collections of documents into classification automatically. It can also create and maintain a hierarchical structure of categories which called taxonomy. Classification by providing a nested hierarchy of queries from general to specific provides the ease of browsing while letting to get more and more precise.

In this system, the classification process is being applied on news text which is the proceeding articles from web. This process is called text classification. News text classification is the process of assigning documents into a fix number of predefined categories. News filtering is one kind on text categorization. Based on Dong and Han, text categorization is an important application and domain of machine learning. It has many special traits such as huge number of features which is hundreds of thousands features to be considered.

News filtering is an important component in many information organization and management task. It is used in many applicative contexts which are in any application requiring document organization or selective.

There are also occur problems in news text filtering. Such of the problems are text classification normally involve an extremely high dimensional space which means that the features usually represent words derived from the textual content of documents.

2.3 MACHINE LEARNING

In this project, machine learning approach is applied to news text filtering. This approach is suitable for news filtering. Based on Bengio, learning means changing in order to be better when a similar situation arrives. It is also an essential of human property. According to Mitchell, machine learning is multidisciplinary field that brings together scientist from artificial intelligence, computational complexity theory, probability and statistics, information theory, philosophy.

- Performance Criteria : It is the process of measuring the quality of the learning output in terms of accuracy.

Machine learning is a learning algorithm and it is focus more on categorization area. The learning algorithm for classification task consists of two components which are a learning component and classification component.

Chapter 3: SYSTEM ANALYSIS & PROTOTYPE DESIGN

3.1 SYSTEM ANALYSIS

This chapter focuses on the system analysis and prototype design for news classification system. The system how to operate and how to flow the data, the system analysis will find. System analysis helps the system to be perfect in execution. The prototype design was based on the object-oriented design using Unified Modeling Language.

3.2 NAÏVE BAYES ALGORITHM

The algorithm is divided into two parts

3.2.1 LEARN NAÏVE BAYES TEXT(Example V)

Example is a set of text documents along with their target values. *V* is the set of all possible target values. This function learns the probability terms $P(w_k|v_j)$, describing the probability that a randomly drawn word from a document in class v_j will be the English word w_k . It also learns the class prior probabilities $P(v_j)$.

1. Collect all words, punctuation, and other tokens that occur in Examples
- Vocabulary← the set of all distinct words and other tokens occurring in any text document from Examples
2. Calculate the required $P(v_j)$ and $P(w_k|v_j)$ probability terms
- For each target value v_j in *V* do
    - $Docs_j$← the subset of documents from Examples for which the target value is vj
    - $P(v_j)$← $|docs_j|$ / |Examples|
    - $Text_j$← a single document created by concatenating all members of docsj
    - n← total number of distinct word positions in Textj
    - For each word Wk in vocabulary
        - N_k←number of times word wk occurs in Textj
        - $P(w_k|v_j) \leftarrow (N_k+1)/n+|Vocabulary|$

3.2.2 CLASSIFY NAÏVE BAYES TEXT(Doc)

Return the estimated target value for the document Doc. Aj denotes the word found in the position within Doc

- Positions← all words positions in Doc that contain tokens found in Vocabulary
- Return V_{nb}, where V_{nb}= for each $v_j \square V \rightarrow$argmax $P(v_j) \Pi P(a_i | v_j)$

3.3 UNIFIED MODEL LANGUAGE (UML)

Unified Modeling Language (UML) is a non-proprietary, third generation modeling and specification language. However, the use of UML is not restricted to model software. The UML also allows developers to specify, visualize, and document the artifacts of an object-oriented software intensive system under development. Besides that, it also represents a compilation of best engineering practices which have proven to be successful in modeling large, complex system, especially at the architectural level.

3.3 THE GENERAL ARCHITECTURE DESIGN

Design is the most important and crucial part of a system development process. In designing a system, it involves the understanding of the studied domain, the application of pertinent scientific and technical knowledge, the creation of various alternatives, and the synthesis and evaluation of proposed alternative solutions. Figure 3.3 shows the general prototype architecture design for news classification

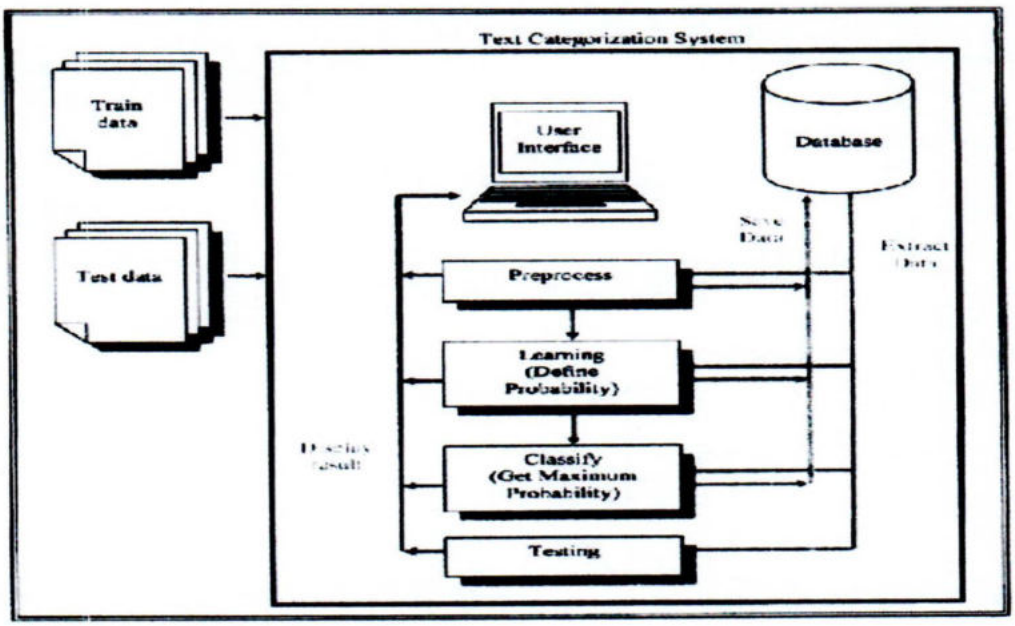

Fig.3.1: The Prototype Architecture Design

In this architectture design, it involved four processes which are preprocess, learning, classify, and testing. All these process interact with database for extracting and saving the data. There are two type of input data for this prototype which are trained data and tested data. These data were saved into the database first before start the preprocess task.

3.4 THE UML DIAGRAM

In this project, the design for the prototype was based on a Unified Modeling Language (UML) that was proposed by Grady Booch. UML is a non-proprietary, third generation modeling and specification language. In addition, it is an open method used to specify, visualize, construct and document the artifacts of an object-oriented software-intensive system under development.

3.4.1 THE USE CASE DIAGRAM

Use cases are description of the functionality of the system from the user's perspective. Use case diagram are used to show the functionality that the system will provide and show which users will communicate with the system in some way to use that functionality.

There are four cases in this prototype for news classification system. These use cases interact with only one actor which is the user for the system. The actor is the system admin. Figure 3.2 shows the use case diagram :

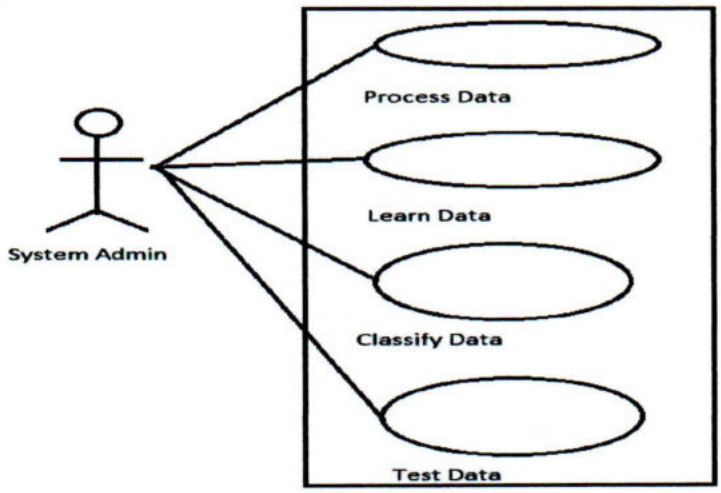

Fig.3.2: Use Case Diagram

- **Preprocess Data** : When user initiaties this use case, the data from the abstract will be extracted in order to obtain the distinct word
- **Learn Data :** This use case for learning data.
- **Classify Data :** This function is to classify the news on the learning function. The system classify the category for each news automatically based on the maximum probability that was obtained from the process.
- **Test Data :** In this use case, the unseen data which is test data was used to validate the accuracy of the prototype system.

3.4.2 THE SEQUENCE DIAGRAM

The Sequence diagram is diagram that shows the object interaction with one another, and these interactions occur over time This sequence shows the time based dynamics of the interaction. Besides that, the sequence shows the time based dynamics of the interaction. Besides that, the sequence diagram can be drawn at different levels of details and to meet different purposes at several stages in the development life cycle.

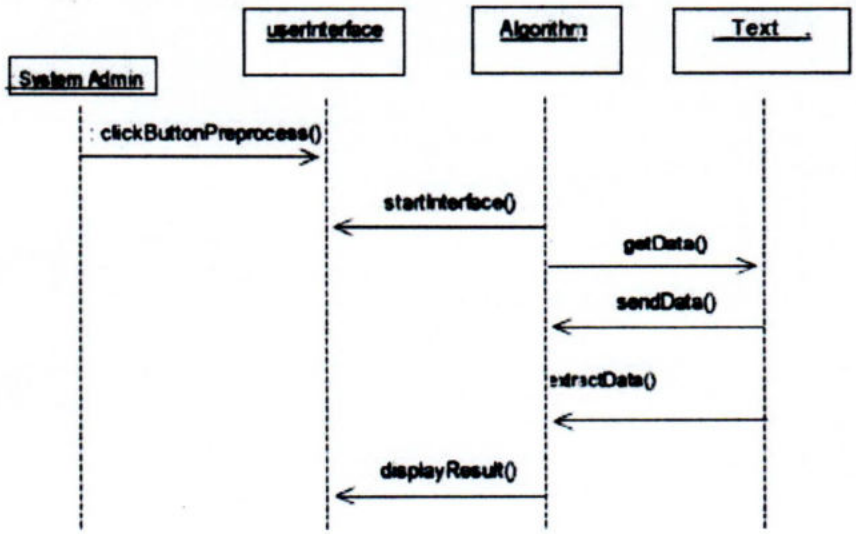

Fig.3.3: Sequence Diagram for Process

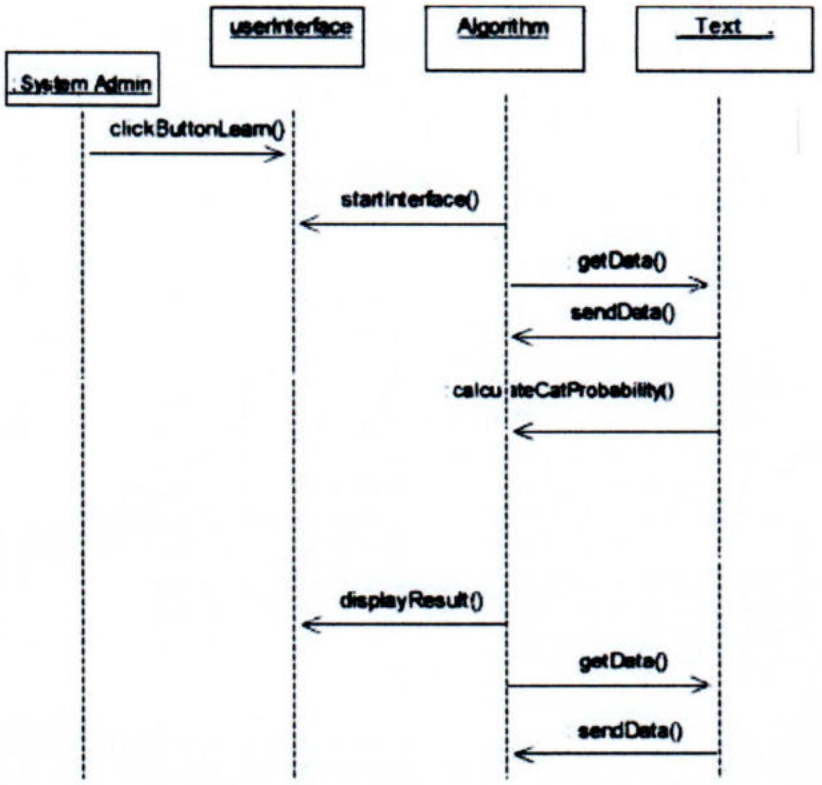

Fig.3.4: Sequence Diagram for Learning

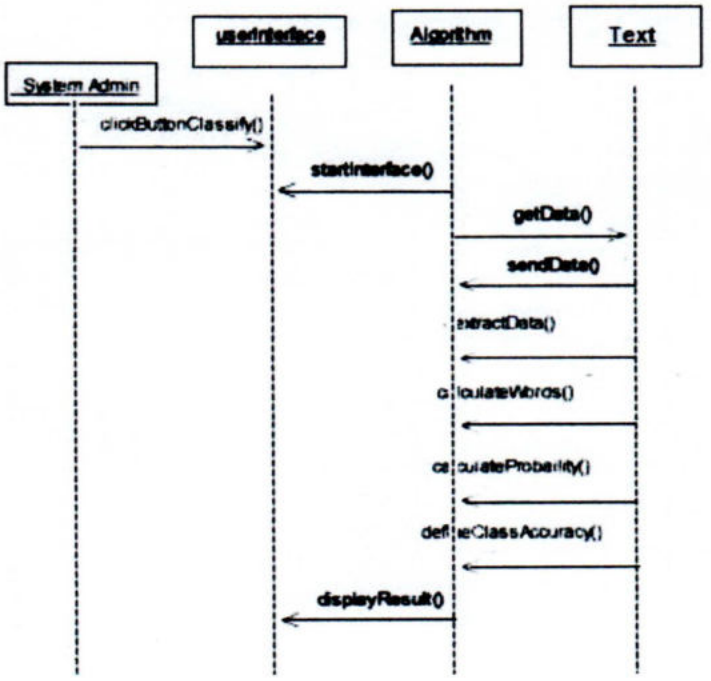

Fig.3.5: Sequence Diagram for Classify

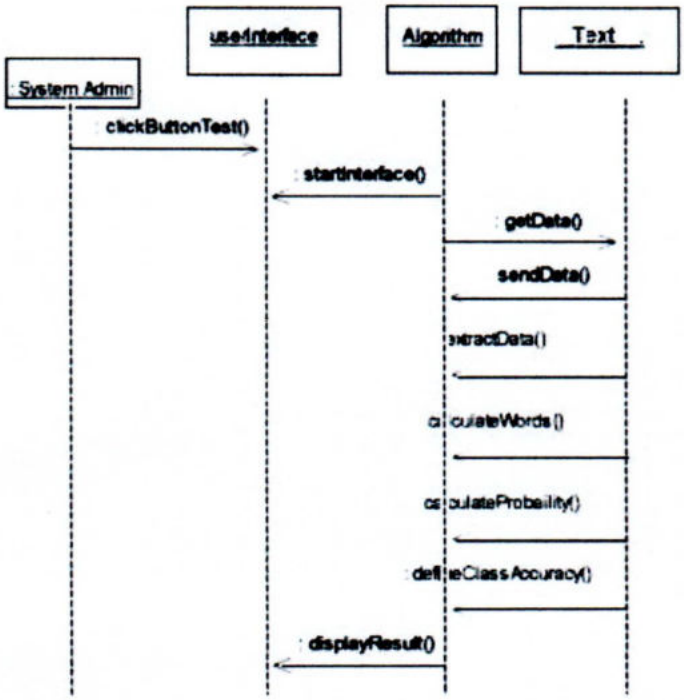

Fig.3.6: Sequence Diagram for Test

Chapter 4: Software Development

4.1 DEVELOPMENT

This chapter focuses on the implementation of the prototype based on the naïve Bayes algorithm. The process of classifying text news using naïve Bayes is explained in details in this chapter.

4.1.1 PHP

PHP version 5 is used in develop this prototype system for this study. PHP is an object-oriented programming language and sometimes called best handler because it enables programmers to quickly build prototype applications. This is a programming language for web. PHP can access data on the HTML pages. Difficult mathematical calculation can be possible using PHP.

4.1.2 MySQL DATABASES

To access and store huge trained data MySQl database is used. This database is user friendly and very much easy to interact with the PHP. MySQL is light database with security of the data.

4.1.3 THE PROCESS AND ALGORITHM FOR TEXT CLASSIFICATION

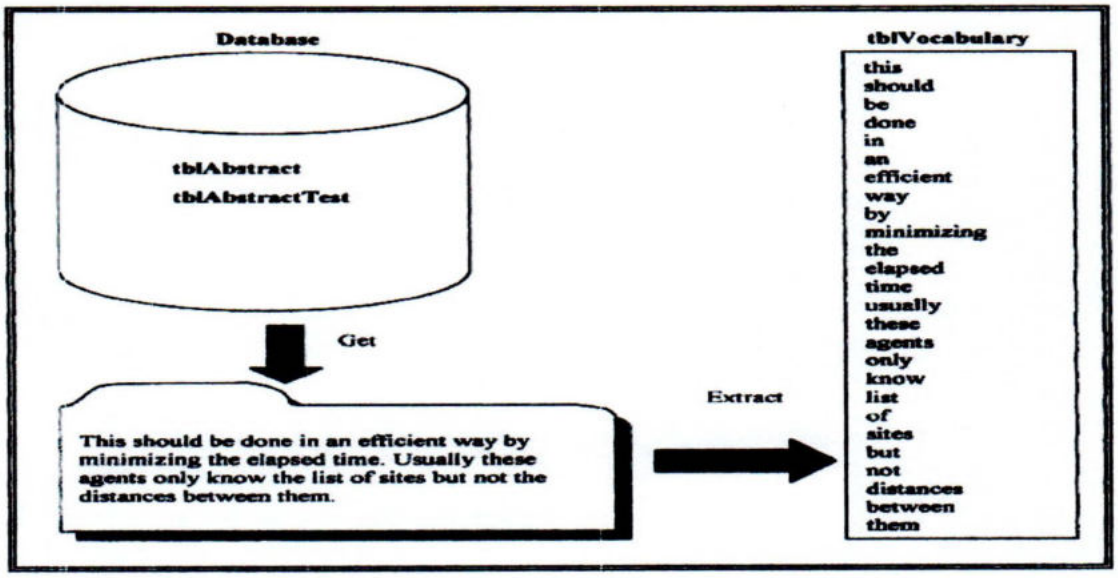

Fig.4.1: Text Processing

The algorithm for learn the train data :

LEARN_NAIVE_BAYES (*Examples*)

Examples is a set of text documents along with their target values which V is the set of all possible target values. This function learns the probability terms $P(w_k \mid v_j)$, and can also learn the class prior probabilities $P(v_j)$.

- Calculate the required $P(v_j)$ and $P(w_k \mid v_j)$ probability terms

  For each target value v_j in V do

  - $docs_j \leftarrow$ the subset of documents from *Examples* for which the target value is v_j
  - $P(v_j) \leftarrow \dfrac{|docs_j|}{|Examples|}$
  - $Text_j \leftarrow$ a single document created by concatenating all members of $docs_j$
  - $n \leftarrow$ total number of distinct word positions in $Text_j$
  - For each word w_k in *Vocabulary*

    $n_k \leftarrow$ number of times word w_k occurs in $Text_j$

    $$P(w_k \mid v_j) \leftarrow \frac{n_k + 1}{n + |Vocabulary|}$$

Fig.4.2. Algorithm for text classification

According to the algorithm there were calculate the prior probability and for each distinct work find the number of occurances.

To find the accurate probability at first finds the total number of words. To find the probability of a particular class, calculates the total number of trained category and the number of particular category. From these two calculation the category's probability is found.

The next process is to find the probability of each distinct word for a particular class or category of news. If no word is found to then at least one word is taken as a number of occurance. After calculating the prior probability and the probability of individual category on a distinct word is multiplied to find the posteriori probability of the

particular news class. To find the actual probability of each category, calculates total number of words for a particular category from the table 'text'. Then to find the each category's probability the summation of each word occurs was found from the table 'text'. For each news category's probability the total number of occurance of a particular word for a particular news type is divided by total number of words in that news category. All these number of occurances are stored after pre-processing the trained instances.

The algorithm for classifying the news categories.

CLASSIFY_NAIVE_BAYES_TEXT (*Text*)

Returns the estimated target value for the document Text. a_i is the word found in the *i*th position within *Text*.

- **positions ← all word positions in Text that contain tokens found in *Vocabulary***
- **Return V_{NB} where**

$$V_{NB} = \arg\max_{v_j \in V} \prod_i P(a_i \mid v_j)$$

Fig.4.3: Algorithm for classify text

The code for this part of algorithm :

```
echo "<br> <br> <br>"." Count by each type = "."<br><br><br><br><br>";

 $queri = "SELECT
ans,COUNT(*),SUM(football),SUM(chip),SUM(election),SUM(hockey),SUM(hardware),SUM(policy),SUM(cpu),SUM(cricket),
SUM(vote) FROM classification group by ans";

    $check = mysql_query($queri);
    $marg = 0;
    $ii = 1;
    $percent =0;
    while($row = mysql_fetch_array($check))
    {
      $attribute = $row['ans'];

      echo " Type = ".$row['ans']; echo "<br>";
      echo " How many Times = ".$row['COUNT(*)']; echo "<br>";

      echo " Sum of Attribute Football = ".$row['SUM(football)']; echo "<br>";
      echo " Sum of Attribute Chip = ".$row['SUM(chip)']; echo "<br>";
      echo " Sum of Attribute Election = ".$row['SUM(election)']; echo "<br>";
      echo " Sum of Attribute Hockey = ".$row['SUM(hockey)']; echo "<br>";
      echo " Sum of Attribute Policy = ".$row['SUM(policy)']; echo "<br>";
      echo " Sum of Attribute Cricket = ".$row['SUM(cricket)']; echo "<br>";
      echo " Sum of Attribute Vote = ".$row['SUM(vote)']; echo "<br>";
      echo " Sum of Attribute CPU = ".$row['SUM(cpu)']; echo "<br>";
      echo " Sum of Attribute Hardware = ".$row['SUM(hardware)']; echo "<br>";

      $val = $row['SUM(football)'];
      $va2 = $row['SUM(chip)'];
```

```
$in5=pow(($va5/$sum),$tt5);
$in6=pow(($va6/$sum),$tt6);
$in7=pow(($va7/$sum),$tt7);
$in8=pow(($va8/$sum),$tt8);
$in9=pow(($va9/$sum),$tt9);

$multi = $in1*$in2*$in3*$in4*$in5*$in6*$in7*$in8*$in9;
$final = $prior*$multi;
$fi = log($final);
echo "<br>";
echo " <br> Total No. of Words = ".$ssum;
echo " <br> Log of (Total Words) = ".$sum;
echo " <br><br> Prior Probability of this type = ".$prior;
echo " <br>     Multi= ".$multi;
echo " <br> Probability for this class (Proior * Multi) = ".$final."<br>";
echo " Probability for this class After Log (Proior * Multi) = ".$fi."<br><br>";

// aa[$ii]=$final;
$ii++;

$percent = $final + $percent;

if($marg < $final)
{
    $marg = $final;
    $win_class = $row['ans'];
}
```

Fig.4.4: Code for algorithm

Chapter 5: RESULT AND DISCUSSIONS

5.1 OVERVIEW

This chapter concentrates on the finding from the experiments on text categorization prototype. In this study there were 60% articles from the web for the training and 40% is used to test the system. There collection of the news type were divided into six different categories. To classify these news articles the Naïve Bayes text classification algorithm was used. The classification for certain article is based on the maximum value of the naïve bayes probability estimation.

5.2 RESULT AND DISCUSSIONS

In this study proposed algorithm had been tested to categorize proceeding articles automatically. The news articles are from different six groups. For training and testing two groups were created. So, total articles were splited into two groups.

In classifying the articles, it must follow the naïve bayes algorithm which divided into three processes. The processes are preprocess, learning, and classifying. After the preprocessor step, the total number of distinct words was obtained. Based on this experiment, there are total 1281 words produced and used for calculation.

The prototype is divided into two categories in order to validate the performance of the algorithm. The classification are result based on training articles and the testing articles. There were 67 articles were experimented as a training news article and classify the document based on the maximum probability that obtained from the classification process. About 80% of the total training and the testing the result is accurate.

The total number test instances were successfully tested. But the system sometimes cause some unwanted result. Due to the huge amount of data the filtering techniques sometimes the calculated result value increases and sometimes decreases. In this filtering logarithm is used to solve this problem. In future techniques this type of mathematical problem must be solved.

REFERENCES

Mitchell, T, M (1997) Machine Learning McGraw Hill

Naïve Bayes Classifier Wikipedia

A. K. McCallum, "Naive Bayes algorithm forlearning to classify text," 1997. http://www-2.cs.cmu.edu/afs/cs/project/theo-11/www/naive-bayes.html , last accessed on April 1st, 2011.

D. Hand, H. Mannila, and P. Smyth, Principles of Data Mining. Cumberland: The MIT Press, 2011.

A. K. McCallum, "Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering." http://www-2.cs.cmu.edu/~mccallum/bow last accessed on December, 2011